

Charles University in Prague
Faculty of Science
Developmental and Cell Biology



Prasoon Kumar Thakur

Decoding RNA splicing regulation and evolution using human long-non-coding RNAs and diplonemids pre-mRNA as models

CZ: Rozluštění regulace a evoluce RNA sestřihu s využitím lidských dlouhých nekódujících RNA a pre-mRNA z diplonemid

Ph.D. Thesis

Supervisor: Prof. David Staněk, PhD.

Department of RNA biology,
Institute of Molecular Genetics AS CR

Prague, 2025

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

Praha, 2025

Prasoon Kumar Thakur

Disclaimer:

I declare that I wrote the thesis independently and that I stated all the information sources and literature. This work or a substantial part of it was not presented to obtain another academic degree or equivalent.

Prague, 18.02.2025

Prasoon Kumar Thakur

Signature

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my supervisor, Prof. David Staněk, for his consistent support, encouragement, and invaluable guidance throughout my PhD program. His mentorship created a welcoming and friendly environment that was essential for my academic and personal growth.

I would like to acknowledge my collaborators, Prof. Julius Lukeš, Dr. Aleš Horák, and Dr. Anzhelika Butenko from the Institute of Parasitology, Biology Centre, CAS, v.v.i., České Budějovice, as well as Prof. Michaela Müller-McNicoll from the Institute for Cell Biology and Neuroscience, Goethe University Frankfurt, Germany, for their valuable contributions to my research. I would like to thank Prof. Carol J. Wilusz from the Department of Microbiology, Immunology & Pathology, Colorado State University, Colorado, USA, for reading this PhD thesis and providing valuable feedback.

I extend my sincere thanks to all current and former members of the RNA Biology Laboratory, particularly Zuzana Cvačková, Filip Karásek, Zuzana Krchnáková, Michaela Krausová, Felix Zimmann, Klára Klimešová, Adriana Roithová, Davide Basello, Nenad Radivojević, Poulami Banik, Mina Obuća, Hana Petržílková and Jana Machatová Křížová, their friendship, assistance, and the joyful moments we shared were instrumental in making this journey memorable. Furthermore, my friend Can Balaban deserves recognition for being a partner in various activities, adding joy and balance to my life outside the lab.

My deepest appreciation goes to my family, especially my parents, grandparents Shiv Kala Devi and Dev Chandra Mishra, my sister Ritu, and my cousins Neha and Abhishek. Their boundless love and support have been my constant source of strength. A special thanks to my elder brother, Prashant Thakur, for his firm belief in me and his continuous encouragement, which have been particularly motivating. I would also like to acknowledge my parents-in-law for their love and encouragement throughout my studies. Lastly, I am profoundly grateful to my wife, Lopita, for her unwavering love, care, and support during this journey. Her presence has been a cornerstone of my perseverance and success.

List of papers and author's contribution

Research articles:

1. Krchnáková Z, **Thakur PK**, Krausová M, Bieberstein N, Haberman N, Müller-McNicoll M, Stanek D. (2019) *Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR proteins*. Nucleic Acids Res.47(2):911-928. doi:10.1093/nar/gky1147.
Prasoon Thakur design the bioinformatics analysis and interpreted data.
2. Zimmann F*, McNicoll F*, **Thakur PK***, Blažíková M, Kubovčíak J, Hernández Cañas MC, Nováková Z, Bařinka C, Kolář M, Staněk D, Müller-McNicoll M, Cvačková Z. (2024) *Retinitis pigmentosa-linked mutations alter the RNA binding of PRPF8 and SNRNP200*. Cellular and Molecular Life Sciences (Accepted).
Prasoon Thakur performed bioinformatics analysis, interpreted data and participated in manuscript writing. His contribution is documented by the fact that he co-shares the first authorship on this publication.
3. **Thakur PK**, Butenko A, Karásek F, Svobodová M, Faktorová D, Pavlisková H, Varga V, Horák A, Lukeš J, Staněk D. (2024) *Splice site diversity and abundance of non-canonical introns in diplomemids (Diplonemea, Euglenozoa)*. RNA (In review).
Prasoon Thakur was the main driving force of the project, design the bioinformatics analysis, interpreted data and wrote the first draft of the manuscript and actively participated during the whole manuscript writing.

Review article:

4. Thind AS, Monga I, **Thakur PK**, Kumari P, Dindhoria K, Krzak M, Ranson M, Ashford B. (2021) *Demystifying emerging bulk RNA-Seq applications: the application and utility of bioinformatic methodology*. Brief Bioinform. 22(6). doi: 10.1093/bib/bbab259.
Prasoon thakur participated in manuscript writing and manuscript editing.

Book chapters:

5. **Thakur PK**, Rawal HC, Obuca M, Kaushik S. (2018) *Bioinformatics Approaches for Studying Alternative Splicing*. In: Reference Module in Life Sciences. DOI: 10.1016/B978-0-12-809633-8.20228-8.
Prasoon Thakur participated in design, manuscript writing, and manuscript editing.
6. **Thakur PK**, Staněk D. (2020) *A review of computational resources for the model organism*. Advances in disease model. (pages 272-284).
Prasoon Thakur participated in design, manuscript writing, and manuscript editing.

* These authors contributed equally

Abstract

RNA splicing is a fundamental process in eukaryotic gene expression, where functional RNA sequences (exons) interrupted by non-coding regions (introns) are joined to produce mature RNA molecules. This process is mediated by the spliceosome, a dynamic macromolecular complex. My thesis investigates two underexplored areas of RNA splicing: the molecular basis for the lower splicing efficiency of long non-coding RNAs (lncRNAs) and the diversity of canonical and non-canonical introns in diplonemid species.

lncRNAs undergo similar maturation steps as protein-coding pre-mRNAs (PCGs), yet they are often spliced less efficiently. Using ENCODE RNA-seq data, we confirmed that lncRNAs exhibit lower splicing indices across multiple human cell lines. Further, our genome-wide analyses revealed that lncRNA splicing efficiency positively correlates with the strength of the 5' splice site (5'ss), a pattern not observed in PCGs. Additionally, efficiently spliced lncRNAs exhibit a higher thymidine content in their polypyrimidine tracts (PPTs) compared to PCGs. lncRNA exons also harbor fewer predicted binding sites for splicing regulatory (SR) proteins, which are critical for efficient spliceosome recruitment. Using iCLIP for SRSF2, SRSF5, and SRSF6, and analyzing eCLIP data for SRSF1, SRSF7, and SRSF9, we demonstrated that SR protein binding to lncRNA exons is markedly reduced compared to expression-matched PCGs. These findings suggest that lncRNAs lack the cooperative splicing enhancer network present in PCGs, rendering their splicing more dependent on the intrinsic strength of splice sites.

In contrast to the conserved canonical intron structures observed in most eukaryotes, euglenozoans, unicellular flagellates that include free-living euglenids, pathogenic kinetoplastids, and marine diplonemids, display remarkable intron diversity. Diplonemids, in particular, feature highly unconventional splicing mechanisms. Our comprehensive analysis of six diplonemid species (*Artemidia motanka*, *Namystynia karyoxenos*, *Sulcionema specki*, *Lacrimia lanifica*, *Diplonema japonicum*, and *Rhynchopus humris*) revealed that all possess a nearly complete set of U2-type spliceosomal small nuclear ribonucleoproteins (snRNPs), indicative of a functional spliceosome. However, in two closely related species, *Artemidia motanka* and *Namystynia karyoxenos*, the majority of introns are non-canonical and lack the conserved GT-AG terminal

dinucleotides typical of U2-type introns. These non-canonical introns exhibit extensive base-pairing potential, suggesting that RNA secondary structure facilitates intron removal by bringing splice sites into close proximity. Despite the conserved splicing machinery across diplomids, splice site diversity varies significantly among species. This work provides insights into the molecular underpinnings of splicing efficiency in lncRNAs and highlights the evolutionary flexibility of splicing mechanisms in eukaryotes.

Abstrakt

Sestřih RNA je základním procesem při expresi eukaryotických genů, při němž se spojují funkční sekvence RNA (exony) přerušené nekódujícími oblastmi (introny) a vznikají maturované molekuly RNA. Tento proces je řízen dynamickým makromolekulárním komplexem, nazývaný spliceosomem. Ve své diplomové práci se zabývám dvěma málo prozkoumanými oblastmi sestřihu RNA: základními molekulárními procesy nižší účinnosti sestřihu dlouhých nekódujících RNA (lncRNA) a rozmanitostí kanonických a nekanonických intronů u diplonemidních druhů.

LncRNA procházejí podobnými maturačními kroky jako pre-mRNA kódující proteiny (PCG), přesto jsou často sestřihávány méně efektivně. Pomocí dat ENCODE RNA-seq jsme potvrdili, že lncRNA vykazují nižší sestřihové indexy napříč několika lidskými buněčnými liniemi. Naše celogenomové analýzy dále odhalily, že účinnost sestřihu lncRNA pozitivně koreluje se silou 5' sestřihového místa (5'ss), což u PCGs nebylo pozorováno. Kromě toho efektivně sestřihané lncRNA vykazují vyšší obsah thymidinu ve svých polypyrimidinových traktech (PPT) ve srovnání s PCG. Exony lncRNA také obsahují méně předpokládaných vazebných míst pro splicingové regulační proteiny (SR), které jsou velmi podstatné pro účinné nasedání spliceosomu. Pomocí iCLIP pro SRSF2, SRSF5 a SRSF6 a analýzou dat eCLIP pro SRSF1, SRSF7 a SRSF9 jsme prokázali, že vazba SR proteinů na exony lncRNA je ve srovnání s expresně odpovídajícími PCG výrazně snížena. Tato zjištění naznačují, že lncRNA postrádají kooperativní síť zesilovačů sestřihu přítomnou u PCG, což způsobuje, že jejich sestřih je více závislý na vnitřní síle sestřihových míst.

Na rozdíl od konzervovaných kanonických struktur intronů pozorovaných u většiny eukaryot vykazují euglenozoa, jednobuněční bičíkovci, mezi něž patří volně žijící euglenidy, patogenní kinetoplastidy a mořské diplonemidy, pozoruhodnou rozmanitost intronů. Zejména diplonemidy se vyznačují velmi nekonvenčními mechanismy sestřihu. Naše komplexní analýza šesti druhů diplonemidů (*Artemidia motanka*, *Namystynia karyoxenos*, *Sulcionema specki*, *Lacrimia lanifica*, *Diplonema japonicum* a *Rhynchopus humris*) odhalila, že všechny mají téměř kompletní sadu malých jaderných ribonukleoproteinů (snRNP) typu U2, což svědčí o funkčním spliceosomu. U dvou blízce příbuzných druhů, *Artemidia motanka* a *Namystynia karyoxenos*, je však většina intronů nekanonická a postrádá konzervované koncové dinukleotidy GT-AG typické pro introny typu U2. Tyto nekanonické introny vykazují rozsáhlý potenciál párování bází, což naznačuje, že

sekundární struktura RNA usnadňuje odstranění intronu tím, že přivádí místa sestřihu do těsné blízkosti. Navzdory konzervovanému sestřihovému mechanismu u diplomemidů se diverzita sestřihových míst u jednotlivých druhů výrazně liší. Tato práce poskytuje vhled do molekulárních základů účinnosti sestřihu u lncRNA a zdůrazňuje evoluční flexibilitu mechanismů sestřihu u eukaryot.

Contents

Glossary	12
1. Summary	15
2. Aims of the Study	17
3. Literature Review.....	18
3.1 Gene expression	18
3.2. RNA splicing.....	19
3.2.1 Types of introns in RNA splicing or cis-splicing	20
3.2.1.1. Group I self-splicing introns	20
3.2.1.2. Group II Self-Splicing Introns	23
3.2.1.3. TransferRNA (tRNA) Introns	24
3.2.1.4. Spliceosomal introns	26
3.2.1.4.1. Regulation of Splicing	33
3.2.1.4.2. Constitutive and alternative splicing in mRNA diversity	33
3.2.1.4.3. Type of alternative splicing	34
3.2.3. Intron origin and evolution	35
3.3. LncRNAs.....	36
3.3.1. Origin and classification of lncRNAs.....	36
3.3.2. Localization and Function	38
3.4. Diplonemids: an overview within euglenozoa	39
3.4.1. Genome size variation in kinetoplastids and euglenids:.....	41
3.4.2. Marine Diplonemids: diversity, distribution	41
3.4.3.1. Canonical spliceosomal introns in euglenids	44
3.4.3.2. Non-canonical introns in euglenids	45
3.4.3.3. Intermediate introns in euglenids: Connecting canonical and non-canonical splicing.....	45
3.4.4. Non-canonical introns in marine diplomemids: insights into their structure, function, and evolution	46
4. Project 1 - MATERIAL AND METHODS	49
Calculation of splice site strength	49
Splicing index, exon/intron length and polypyrimidine tract analysis.....	49

Prediction of secondary structures	50
Srsf-binding and splicing silencer motif prediction	50
iCLIP library preparation	51
Analysis of iCLIP and RNA-seq data	51
Mapping of eCLIP sequence data	52
5. Project 2 - MATERIALS AND METHODS	54
6. Results - Project 1	59
5' ss sequence is important for lincRNA splicing	59
Polypyrimidine tract sequence determines the splicing efficiency of lincRNAs	63
SR proteins bind less efficiently to lincRNAs	68
7. Results - Project 2	76
Canonical and non-canonical introns in diplonemid genomes	76
Characterization of the spliceosome components in diplonemids	81
Splice site analysis by comparative genomics	93
8. Discussion	95
Roles of the 5' splice site, Polypyrimidine tract, and splicing factors in lincRNA splicing	95
Diplonemid introns: diverse, non-canonical, and evolutionarily intriguing	96
Integration of findings: evolutionary and mechanistic parallels in splicing adaptations	98
9. Conclusions	100
References:	101

Glossary

3'ss	3' splice site
5'ss	5' splice site
5mC	5-methyldeoxycytosine pathway
AS	Alternative splicing
ATP6V1C1	V-type proton ATPase subunit C1
BACs	Bacterial artificial chromosomes
BP	Branch point
BPP	Base-Pairing Probability
BPS	Branch point sequences
CDC48	Cell Division Control protein 48
CRISP	Cysteine-Rich Secretory Protein
CS	Constitutive splicing
ctDNA	Chloroplast DNA
DESI1	Desumoylating Isopeptidase 1
DI	Domain I
DIV	Domain IV
DSPDs	deep sea pelagic diplonemids
DV	Domain V
DVI	Domain VI
EndA	Endonucleases
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
iCLIP	individual-nucleotide resolution cross-linking and immunoprecipitation
eCLIP	Enhanced CLIP
FLC	Full-length circular intron RNA
GK	Glycerol Kinase
GTP	Guanosine-5'-triphosphate
hnRNPs	Heterogeneous nuclear ribonucleoprotein proteins

IGV	Integrated Genome Viewer
IQR	Interquartile range
ISEs	Intronic splicing enhancers
ISSs	Intronic splicing silencers
KMP-11	Kinetoplastid membrane protein-11
LECA	Last eukaryotic common ancestor
lincRNAs	Long intergenic non-coding RNAs
lncRNA	Long non-coding RNA
MES	Maxent score
MITEs	Miniature inverted-repeat transposable elements
ORF	Open reading frame
OTUs	Operational taxonomic units
PCGs	Protein-coding genes
POC1	Poc1 centriolar protein
PPT	polypyrimidine tract
pre-mRNA	Precursor RNA transcript
pre-tRNA	Precursor tRNA
PWMs	Position Weight Matrices
RiboMinus	rRNA depletion
RNAi	RNA interference pathway
RPE	Retinal pigment epithelial
RPKM	Reads Per Kilobase of transcript per Million
RPL37e	Ribosomal Protein L37e
RPL6	Ribosomal Protein L6
RT	Reverse transcriptase
SCG	Single-cell genomic
SI	Splicing index
SL	Spliced-leader
SNF7	Sorting-associated protein 32
SL <i>trans</i>	Spliced-leader <i>trans</i> -splicing
snRNAs	Small nuclear RNAs

SR	Splicing regulatory
SRSF	Serine and arginine-rich splicing factor
SS	Splice site
TE	Transposable element
tRNAs	Transfer RNAs
U2AF	U2 auxiliary factor

1. Summary

RNA splicing is a fundamental biological process where introns are excised from precursor RNAs, and exons are joined to produce mature RNA molecules. This thesis investigates two underexplored aspects of splicing: the splicing inefficiency of long non-coding RNAs (lncRNAs) and the diversity of canonical and non-canonical introns in diplomonids.

In chapter 6, we have investigated the splicing mechanisms of lncRNAs. Recent studies indicate that nascent lncRNAs undergo similar maturation processes as pre-mRNAs of protein-coding genes (PCGs), yet they exhibit significantly lower splicing efficiency. Using ENCODE RNA-seq data, we confirmed that lncRNAs exhibit lower splicing indices across multiple human cell lines. Further, to understand this phenomenon, we analyzed potential splicing inhibitory sequences. Our genome-wide analysis of intergenic lncRNAs (lincRNAs) revealed that these RNAs do not contain a higher prevalence of splicing inhibitors compared to PCGs, suggesting that inhibitory sequences are not the primary reason for inefficient splicing in lincRNAs. Additionally, we found that lincRNA exons have fewer predicted SR protein binding sites and exhibit lower SR protein binding levels compared to expression-matched PCGs. These findings suggest that lincRNAs lack the cooperative splicing enhancer network present in PCGs, making their splicing outcomes more reliant on the intrinsic strength and optimality of their splice sites.

Diplomonids are unicellular marine protists known for their unique biology. In chapter 7, we explored the diversity of introns in six diplomonid species namely *Artemidia motanka* and *Namystynia karyoxenos* (family Hemistasiidae), and *Sulcionema specki*, *Lacrimia lanifica*, *Diplonema japonicum*, and *Rhynchopus humris* (family Diplomonidae). Mapping RNA-seq reads on the draft genome assembly allowed the identification and basic characterization of their introns. Our genome wide analysis showed that while canonical introns are dominant in most diplomonids, non-canonical introns are prevalent in the hemistasiid clade, such as *Artemidia motanka* and *Namystynia karyoxenos*. These non-canonical introns lack the typical GT-AG dinucleotides and instead exhibit extensive base-pairing, suggesting non-conventional splicing mechanisms. Despite these differences, all diplomonids analyzed possess a functional spliceosome, with species like *S. specki* showing adaptations such as an alternative U1 snRNA for recognizing degenerate splice sites. In contrast, *A. motanka* and *N. karyoxenos* likely rely on RNA secondary structures or specialized pathways for splicing their non-canonical introns. Diplomonids have intron-rich

genomes, which exhibit remarkable evolutionary plasticity, highlight their genetic adaptability compared to their nearly intron-free kinetoplastid relatives, resembling the complexity of multicellular organisms.

2. Aims of the Study

The primary objective of this research was to elucidate diversity and evolution of pre-mRNA introns in eukaryotes. The study focuses on two main areas, Project 1.) Uncovering the molecular mechanisms that contribute to the generally lower splicing efficiency of lncRNAs compared to protein-coding pre-mRNAs. Project 2.) Characterizing the introns within diplonemid genomes and investigating the evolutionary significance of non-canonical introns in these species

The specific aims of the thesis include:

- Defining the roles of the 5' splice site (5'ss) and the polypyrimidine tract (PPT) in the splicing of lncRNAs.
- Evaluating the impact of sequence(s) and various splicing factors on splicing of lncRNAs.
- Identifying and characterizing canonical and non-canonical introns in diplonemid genomes.
- Investigating the potential splicing mechanisms at non-canonical splice sites.
- Tracing the potential origins of non-canonical introns by comparing the evolutionary patterns of canonical and non-canonical splice sites across diplonemids.

3. Literature Review

3.1 Gene expression

In eukaryotes, gene expression is an essential biological process that converts DNA-encoded genetic information into functional products. This intricate process begins with transcription, during which DNA is transcribed into precursor messenger RNA (pre-mRNA) within the nucleus. The pre-mRNA undergoes several critical modifications, collectively known as RNA processing. These modifications include the addition of a 5' cap, splicing to remove introns and join exons and the addition of a poly-A tail at the 3' end (**Figure 1**).

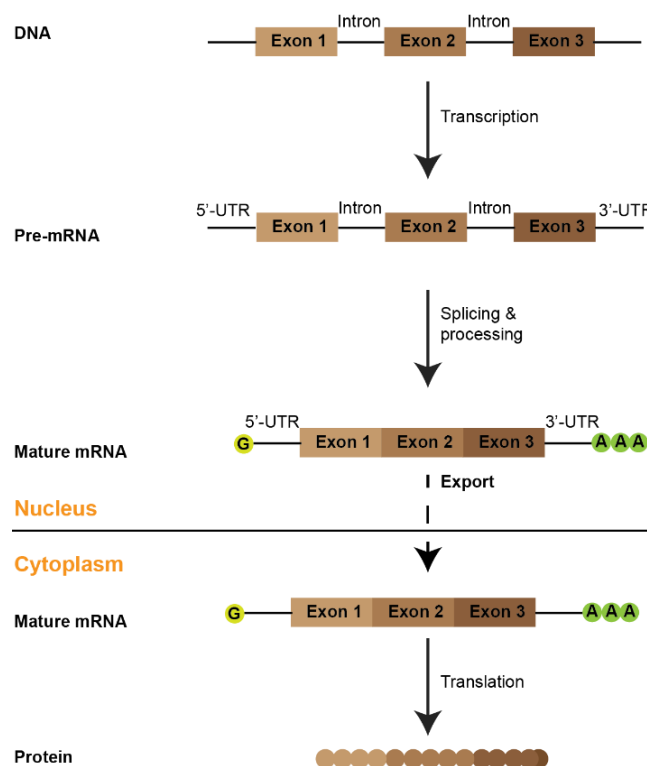


Figure 1. Overview of gene expression.

These steps ensure the mature mRNA's stability, export, and translational efficiency. Once processed, the mRNA is transported to the cytoplasm for translation into a protein. This regulation of gene expression is vital for accurate cellular function and adaptability, with any errors

potentially leading to various diseases, underlining its importance in maintaining cellular health and function (Orphanides and Reinberg 2002).

3.2. RNA splicing

In the 1970s, Phillip Sharp and Richard Roberts (S M Berget et al. 1977; Konkel et al. 1978) made a groundbreaking discovery that initial RNA transcripts (pre-mRNA) in eukaryotes contain non-coding sequences called introns, which must be excised to produce mature mRNA. This discovery overturned the prevailing belief that genes consisted solely of continuous coding sequences in DNA. It also emphasized the importance of RNA processing in gene regulation, influencing our approach to studying genetic information and its utilization within cells (Wilkinson et al. 2020). The splicing mechanism involves transesterification reactions which occur in two sequential steps. First, the 2'-OH group of a branch-point nucleotide within the intron, performs a nucleophilic attack on the first nucleotide of the intron at the 5' splice site (5'ss), forming a lariat intermediate. Second, the 3'-OH group of the released 5' exon attacks the final nucleotide of the intron at the 3' splice site (3'ss), joining the exons together and releasing the intron lariat (**Figure 2**) (Trowitzsch et al. 2009; Hoskins and Moore 2012).

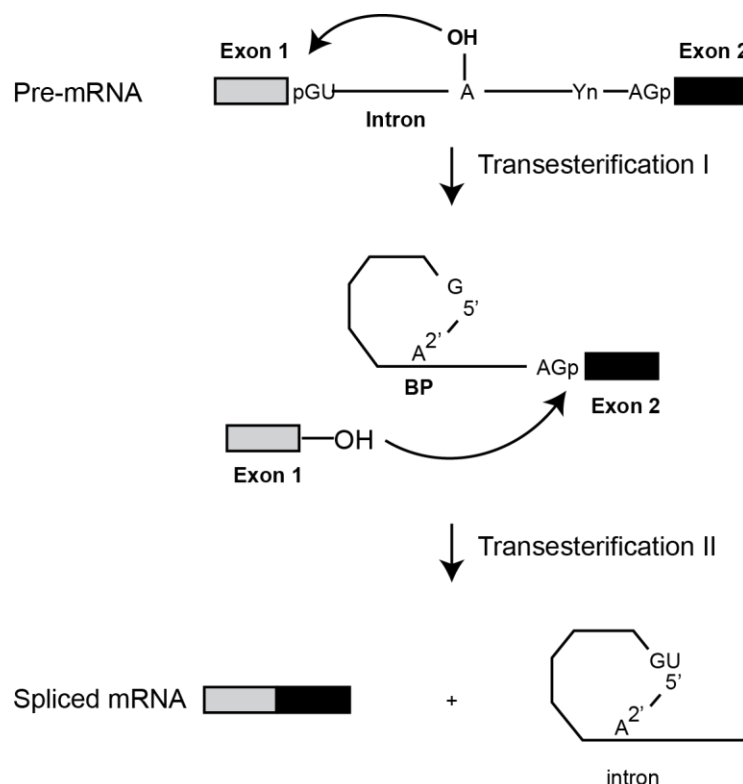


Figure 2. Illustration of the biochemistry steps of splicing.

3.2.1 Types of introns in RNA splicing or cis-splicing

Introns can be classified into four main categories based on their tertiary structure and the mechanism of their excision, 1.) Group I self-splicing introns 2.) Group II self-splicing introns, 3.) tRNA introns (also known as archaeal or BHB introns) and 4) Spliceosomal introns. Introns of self-splicing RNA molecules can splice themselves. Group I and Group II self-splicing introns are ribozymes, meaning they have intrinsic catalytic activity and can remove themselves from the RNA transcript while facilitating exon ligation without the help of protein. tRNA intron splicing is catalyzed by several RNA processing enzymes, which assist in the removal of introns and the subsequent ligation of exons. Spliceosomal introns are excised through the action of a large ribonucleoprotein complex known as the spliceosome, which mediates both intron removal and exon ligation. In the following sections, each intron class is discussed in greater detail.

3.2.1.1. *Group I self-splicing introns*

Group I introns are found in rRNA, tRNA, and protein-coding genes, with a notable abundance in mitochondrial DNA (mtDNA) of fungi and plants (Palmer and Logsdon 1991). Specifically, they are present in nuclear rRNA genes of organisms like *Tetrahymena*, in chloroplast DNA (ctDNA), bacteriophages, and several group I introns have also been discovered in tRNA genes in eubacteria (Michel and Westhof 1990; Palmer and Logsdon 1991; Reinhold-Hurek and Shub 1992). Group I introns are excised through a splicing mechanism involving two transesterification reactions (Cech 1990). First, the 3'-OH group of a guanosine cofactor attacks the 5' splice site. Then, the free 3'-OH of the upstream exon attacks the 3' splice site, joining the exons and releasing the linear intron molecule (**Figure 3**) (Cech 1990). The cofactor can be guanosine or any of its phosphorylated forms for splicing *in vitro*. *In vivo*, it is believed that GTP acts as a cofactor, but this has not been yet confirmed experimentally. Another alternative group I intron processing pathway has been described in literature, in which, intron catalyzed hydrolysis at 3' splice site followed by attack at the 5' splice site by the intron 3' terminal guanosine. As a result, a full-length circular intron RNA (FLC) is formed. The mechanism implies the exons remain un-ligated and presumably leads to a non-functional rRNA product (Nielsen et al. 2003; Haugen et al. 2004).

The GIR1 branching ribozyme, (referred as the lariat capping ribozyme), is a specialized RNA molecule that catalyzes a branching reaction in RNA processing. It is composed of around 180 nucleotides and is part of a unique class of group I introns, specifically within what are known as twin-ribozyme introns. In this process, the 2'-OH group of an internal uridine (U) residue attacks a nearby phosphodiester bond, resulting in the release of a 5' fragment with a 3' 3'-OH and a 3' fragment with the first and third nucleotides connected by a 2',5' phosphodiester bond (Nielsen et al. 2005). This creates a 3-nucleotide lariat structure known as the lariat cap, which likely functions as a substitute for the typical cap found at the 5' end of an intron-encoded mRNA. GIR1 was initially identified as Group I-like Ribozyme I, found alongside GIR2 in what are known as twin-ribozyme introns (Johansen and Vogt 1994; Decatur et al. 1995). Structural studies of GIR1 from the slime mold *Didymium iridis* and various species of the amoeba-flagellate *Naegleria* have revealed a notable resemblance to group I introns splicing mechanism. This similarity is particularly evident with eubacterial tRNA introns, with which GIR1 shares significant sequence similarity and short peripheral elements (Johansen et al. 2002; Nielsen and Johansen 2007).

Group 1 introns exhibit a highly variable distribution, even among closely related species and it has been suggested that this likely due to their mobility as genetic elements. Moreover, there is numerous evidences to show that group I introns are mobile elements and intron mobility at the DNA level by homing is well documented (Dujon 1989; Edgell 2009). Mobility at the RNA level has been suggested to occur by reverse splicing followed by reverse transcription and integration into DNA at cognate (homing) or non-cognate sites (transposition). There is evidence for reverse splicing *in vitro* as well as *in vivo* in *E. coli* in experiments based on the *Tetrahymena* intron (Roman and Woodson 1998) and an intron from the myxomycete *Didymium* (Birgisdottir and Johansen 2005). Circular forms of group I introns could likewise be candidates for intron mobility. Circularity offers resistance towards exonucleases and creates a junction sequence that could be used in recognition of the cognate site in the target RNA. While reverse splicing itself is known to be possible within living organisms (mentioned above), we still lack an experimental system that allows them to fully explore how reverse splicing is stabilized in the genome. Specifically, they have not yet developed ways to observe how the intron, once reverse spliced into an RNA, becomes a permanent part of the DNA through reverse transcription (the process of converting RNA back into DNA) and recombination (the process by which the new DNA sequence is integrated into the genome).

One of the intriguing questions that still remains is the origin of group I introns. The discovery of catalytic RNA, which combines both template and catalytic properties, has led to the hypothesis of an ancient RNA world predating DNA and protein. Group I introns are thought to have given rise to the GIR1 branching ribozyme, which chemically performs the same reaction as the first step of spliceosomal and group II intron splicing. This observation could offer clues about the evolution of spliceosomal splicing (Nielsen and Johansen 2009). Interestingly, group I introns have been identified in tRNA genes in chloroplast DNA (ctDNA) and in five different species of cyanobacteria, suggesting that their existence predates the evolution of plastids. This makes these introns potentially 1 to 3.5 billion years old, aligning with the hypothesis that group I introns are remnants of a primordial RNA world (Saldanha et al. 1993).

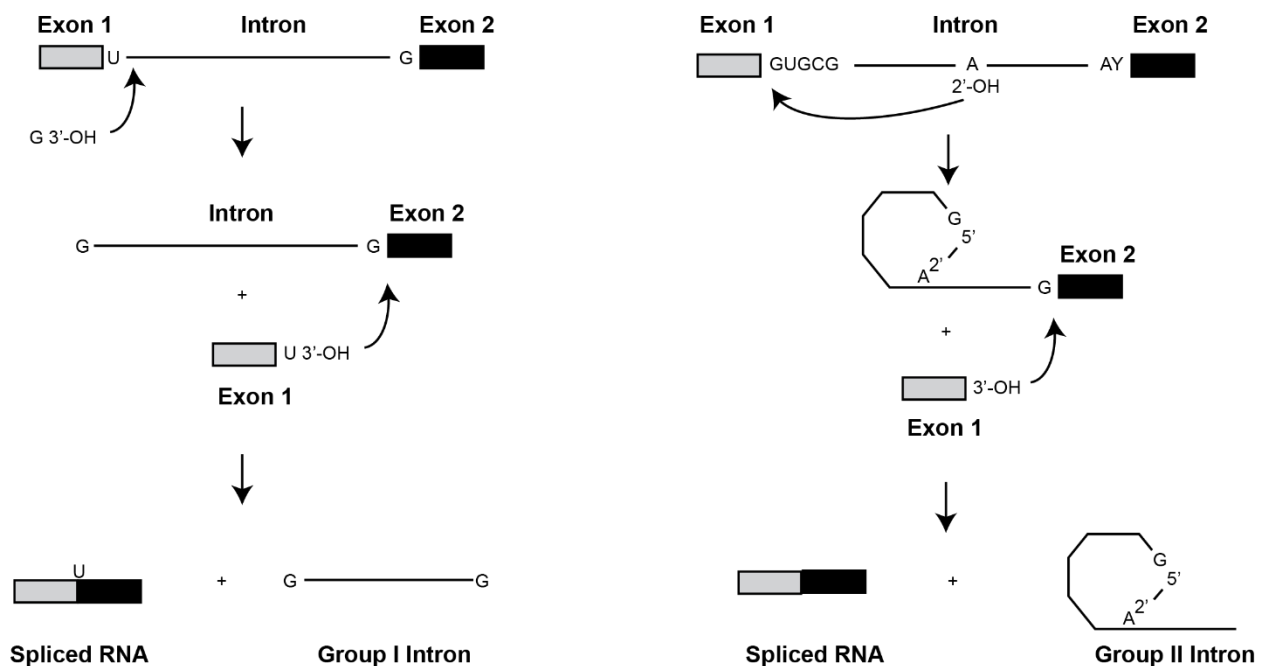


Figure 3. Splicing mechanisms of group I and group II introns. In both cases, a series of transesterification reactions are used to excise the intron and ligate the exons. The net number of bonds remains the same throughout. The reaction is initiated by a guanosine cofactor in group I introns and by an internal adenosine in group II. *Created by author inspired from (Saldanha et al. 1993).*

3.2.1.2. Group II Self-Splicing Introns

Group II introns are highly versatile ribozymes capable of self-splicing and acting as retro transposable elements. They are widely found in bacteria and in bacterial-derived organelles, such as mitochondria and chloroplasts, in fungi, algae, plants, and some early-diverging metazoans. (Zimmerly and Semper 2015). The splicing process begins with the formation of a lariat structure, where the 2'-OH of an internal adenine residue within the intron attacks the 5'ss, creating a lariat with a 2'-5' phosphodiester bond. Next, the 3'-OH of the upstream exon attacks the 3'ss, joining the exons and releasing the intron as a lariat (**Figure 3**). Group II introns have conserved sequences at their 5'- and 3'-boundaries (GUGYG and AY), which bear some resemblance to those in nuclear mRNA introns (Michel et al. 1989). Moreover, most group II intron RNAs are composed of two basic elements: a self-splicing ribozyme and an open reading frame (ORF) encoding a reverse transcriptase (RT) enzyme. The ribozyme is organized into six distinct structural domains. Domain I (DI), the largest, harbors recognition sequences crucial for binding the 5' and 3' exons as well as the branch site nucleophile. The 5' exon interacts with two binding sites in DI, known as exon binding sites, EBS1 and EBS2 (Jacquier and Michel 1987), while the 3' ss is determined by a brief pairing between EBS3 and the 3' exon (Costa et al. 2000). Domains II and III (DII and DIII) contribute to increasing the catalytic efficiency of the splicing process (Fedorova et al. 2003), with the linker J2/3 connecting them and forming a critical part of the active site (Mikheeva et al. 2000; de Lencastre and Pyle 2008). Domain IV (DIV) houses the ORF for reverse transcriptase and includes structural motifs responsible for binding the RT protein (Wank et al. 1999). The most conserved structural component of group II introns is domain V (DV), a short hairpin that features a bulge essential for catalysis (Schmidt et al. 1996). This bulge is located near a "catalytic triad" of highly conserved nucleotides (typically AGC, though CGC is also common) at the base of DV (Boulanger et al. 1995). Lastly, domain VI (DVI) contains an adenosine that initiates the first step of splicing by attacking the 5' ss, leading to the formation of lariat RNA (Pyle and Lambowitz 2006).

Similar to group I introns, the varied distribution of group II introns indicates a recent spread as mobile elements. In eukaryotes, nuclear introns (spliceosomal introns) are removed from pre-mRNA by the spliceosome, a complex made of proteins and small nuclear RNAs (snRNAs). Spliceosomal introns are also removed via a lariat-forming mechanism, which is remarkably

similar to the self-splicing seen in group II introns. It has been widely hypothesized that group II introns are the evolutionary ancestors of spliceosomal introns (Sharp 1985; Cech 1986; Peebles et al. 1986; Sharp 1991; Lambowitz and Belfort 2015). These introns are believed to be extremely ancient, likely originating in bacteria during the primordial "RNA world" (Doolittle 2013). While the exact sequence of events leading to the evolution of mobile group II introns remains unclear, at some point, they acquired reverse transcriptase (RT) and endonuclease functions, allowing them to spread and proliferate within bacterial genomes (Lambowitz and Belfort 2015). It is also possible that RNA self-splicing emerged from an existing retroelement as a means of reducing the damage caused by transposons (Curcio and Belfort 1996; Lambowitz and Belfort 2015).

Group II introns were likely incorporated into eukaryotic genomes during the endosymbiotic event when bacteria were engulfed by eukaryotic cells and evolved into mitochondria and chloroplasts (Palmer and Logsdon 1991; Irimia and Roy 2014; Lambowitz and Belfort 2015). This theory is supported by the presence of group II introns in the bacterial species believed to have evolved into these organelles (Robart and Zimmerly 2005; Lambowitz and Belfort 2015). Spliceosomal introns were found in early eukaryotes, and studies suggest that a primitive spliceosome existed in the last common ancestor of modern eukaryotes (Collins and Penny 2005; Lambowitz and Belfort 2015). One intriguing hypothesis is that the proliferation of introns in early eukaryotes may have driven the evolution of the nuclear envelope, which segregated intron splicing from the process of translation (Martin and Koonin 2006b; Lambowitz and Belfort 2015). This separation would have necessitated the development of splicing machinery capable of functioning in trans, allowing it to splice multiple different introns scattered throughout the genome (Martin and Koonin 2006b; Lambowitz and Belfort 2015).

3.2.1.3. TransferRNA (tRNA) Introns

Transfer RNAs (tRNAs) gene serves as a critical adaptor in the translation of mRNA into proteins. Their maturation involves several essential steps, including the removal of 5'-leader and 3'-trailer sequences, excision of introns, post-transcriptional modifications, and the addition of a 3'-CCA sequence. During tRNA maturation, tRNA splicing is one of the most important steps as it includes excision of introns and ligation of the two halves of exons in the precursor (pre)-tRNA, enabling proper folding and function in protein synthesis. tRNA introns are found in tRNA genes of all

three domains of life, including bacteria, archaea, and eukarya. In these three domains of life, pre-tRNA splicing follows distinct pathways. In certain bacteria and the plastids of higher eukaryotes, group I introns in pre-tRNA are excised through an autocatalytic reaction, which requires an external guanosine-5'-triphosphate (GTP) cofactor (Xu et al. 1990; Haugen et al. 2005). Conversely, in the cytoplasm of eukaryotes and archaea, introns are enzymatically removed by tRNA-specific RNA-splicing endonucleases (EndA)(Abelson et al. 1998), followed by the ligation of exon halves by tRNA ligases (Phizicky et al. 1986; Westaway et al. 1988; Englert et al. 2011; Popow et al. 2011). Specifically, tRNA splicing EndA recognizes and excises the intron, creating two separate halves of the tRNA. Then, a tRNA ligase enzyme joins the two halves by phosphorylating and sealing the RNA backbone, resulting in a mature tRNA molecule (**Figure 4**)(Gerber et al. 2022). In archaea, they are also known as BHB (bulge-helix-bulge) introns due to present of BHB structural motif at the exon–intron boundaries within the anticodon loop of archaeal pre-tRNAs consists of a 4-bp helix flanked on each 3' side with 3 nt in a bulge, which is recognized by a archaeal splicing machinery (Kuhse et al. 1990; Reinhold-Hurek and Shub 1992; Biniszkiwicz et al. 1994). In eukaryotes, EndA has been extensively studied, especially in yeast, *Xenopus*, and humans. It consists of four subunits, named Sen2, Sen15, Sen34, and Sen54 (or α , β , γ , δ) (Rauhut et al. 1990; Trotta et al. 1997; Paushkin et al. 2004; Trotta et al. 2006). In archaea, EndA classified into various types based on subunit composition such as α_4 , α'_2 , $(\alpha\beta)_2$ (Tocchini-Valentini et al. 2005) and more recently identified ϵ_2 type (Fujishima et al. 2011; Hirata et al. 2012). In eukaryotes pre-tRNAs, catalytic mechanism of EndA in tRNA intron excision has been elucidated (Reyes and Abelson 1988; Baldi et al. 1992; Bufardecì et al. 1993). In addition, the general mechanism of pre-tRNA recognition and cleavage by archaeal EndA has also been described (Li et al. 1998; Li and Abelson 2000; Xue et al. 2006). These studies have highlighted an evolutionary connection between archaeal and eukaryotic EndAs, which likely share a common ancestral origin.

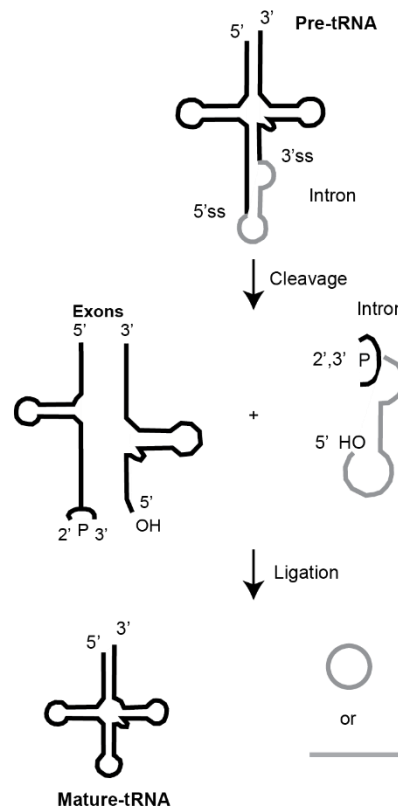


Figure 4. The tRNA splicing pathway. *Created by author inspired from (Zhang et al. 2023).*

3.2.1.4. Spliceosomal introns

Spliceosomal introns are the most common type found in eukaryotic nuclear genes. The process of splicing, which involves the removal of these introns, is facilitated by the spliceosome, which comprises of U1, U2, U5, and U4/U6 (small nuclear ribonucleoproteins) snRNPs and numerous non-snRNP proteins. Each snRNP consists of a small nuclear RNA (snRNA) and a set of common and specific proteins (**Figure 5**) (Trowitzsch et al. 2009; Hoskins and Moore 2012).

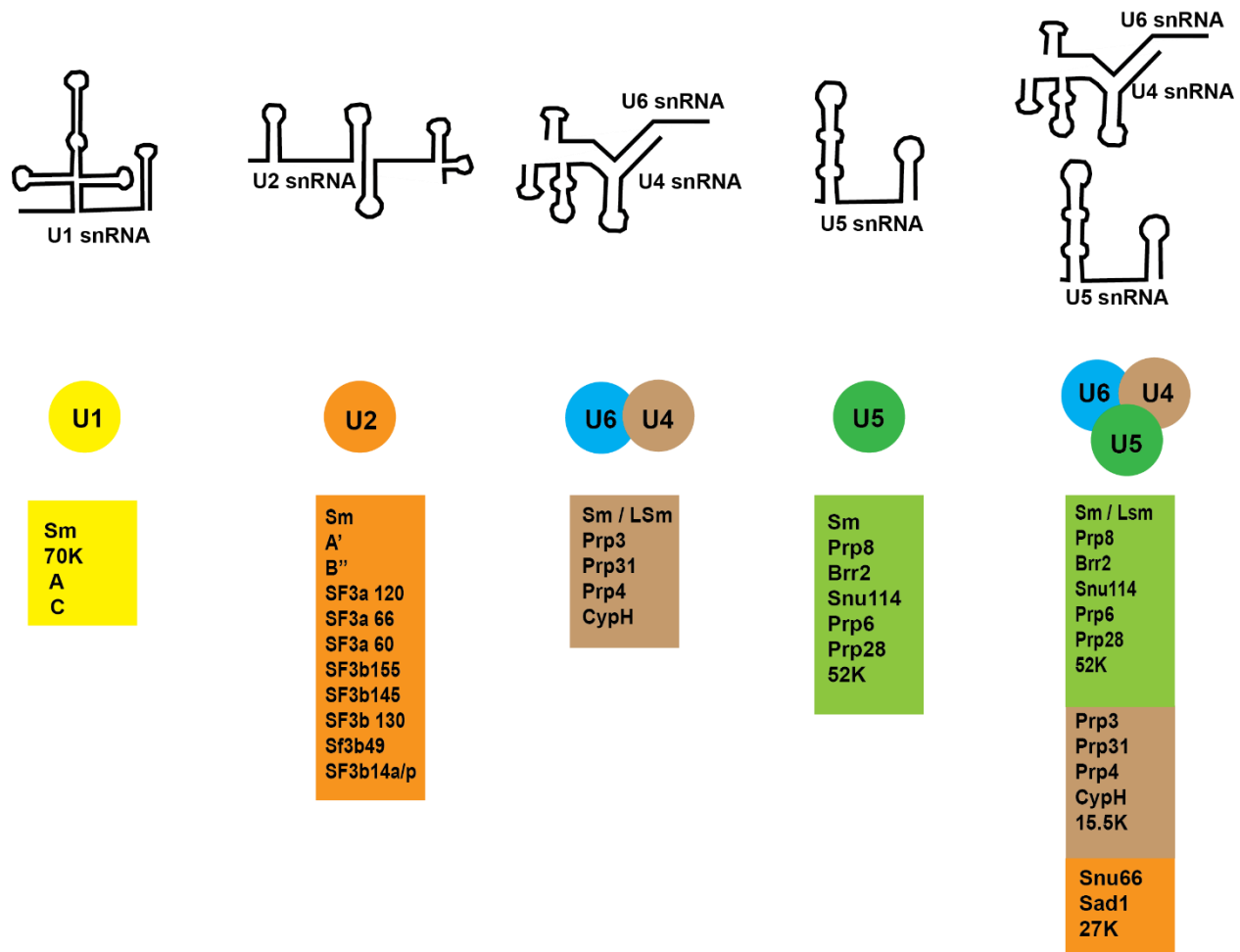


Figure 5: Protein composition and snRNA secondary structures of the human spliceosomal snRNPs. All seven Sm proteins (B/B', D3, D2, D1, E, F, and G) or LSm proteins (Lsm2-8) are indicated by "Sm" or "LSm" at the top showing the proteins associated with each snRNP. The U4/U6.U5 tri-snRNP contains two sets of Sm proteins and one set of LSm proteins. *Created by the author inspired from (Will and Lührmann 2011).*

The spliceosome identifies the boundaries between introns and exons. Key regions involved in splicing include: (a) GT sequences at the start of introns, known as the 5' splice-site (5'ss or donor site); (b) AG dinucleotides at the end of introns, referred to as the 3' splice-site (3'ss or acceptor site); (c) branch point sequences (BPS) located upstream of the 3'ss, which lack high conservation; and (d) polypyrimidine tract (PPT) situated between the BPS and the 3' ss (Zhang et al. 2021). The splicing signal motif (5'ss, 3'ss, and BPS) is highly conserved across eukaryotes (**Figure 6**) (Iwata and Gotoh 2011).

There are two main types of spliceosomes: the major spliceosome (Staley and Guthrie 1998) and the minor spliceosome (Patel and Steitz 2003; Turunen et al. 2013). The major or

canonical spliceosome is the most common and is responsible for splicing the majority of pre-mRNA introns known as U2-type introns. It consists of five snRNPs—U1, U2, U4, U5, and U6—and recognizes canonical splice sites with specific consensus sequences, such as the GT dinucleotide at the 5' ss and the AG dinucleotide at the 3' ss (**Figure 7**).

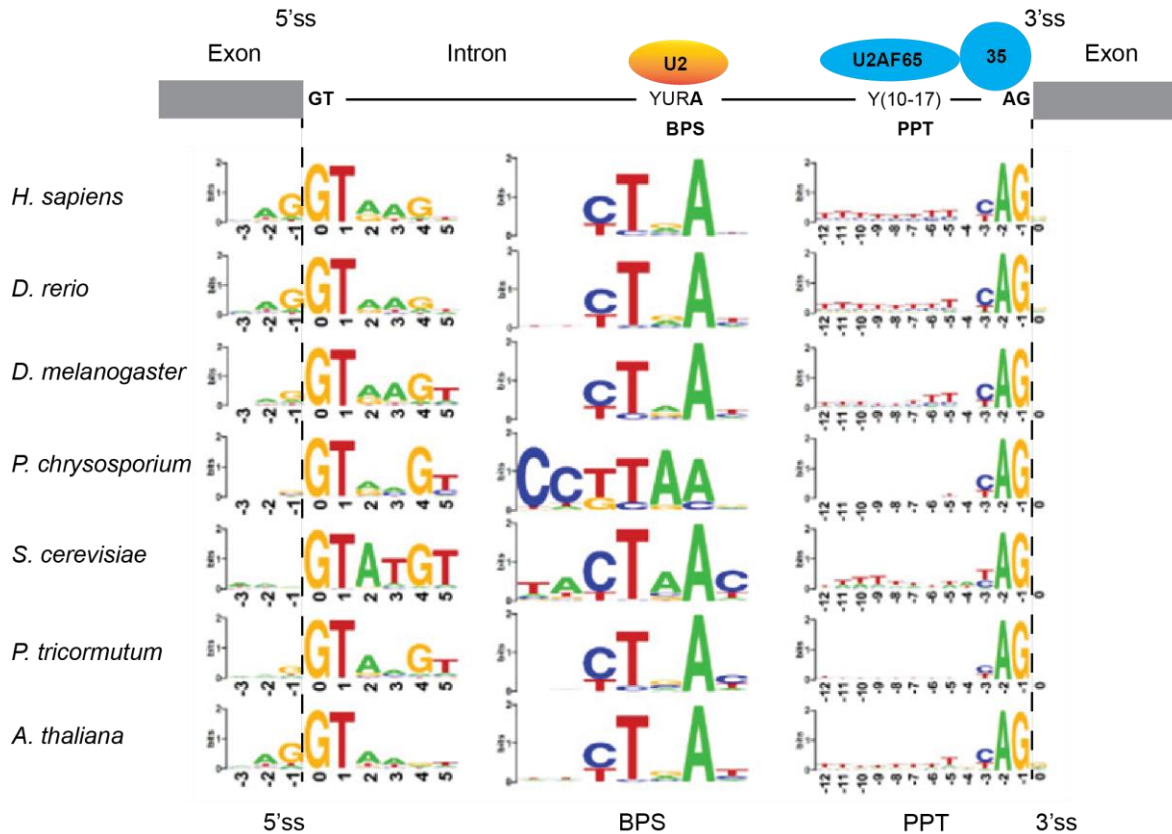


Figure 6: In top, exons are represented by boxes and introns by lines. The most conserved nucleotides at the 5'ss, BPS, PPT, and 3'ss are indicated. In bottom, splicing signal motifs of seven species. Sequence motifs for 5'ss, 3'ss, and BP are depicted by Sequence logos WebLogo. The relative height of each letter is proportional to the relative entropy of the corresponding base at the given position, and bases are listed in descending order of frequency from top to bottom. Splice sites/consensus splice sites are depicted at DNA level. *Adopted from (Iwata and Gotoh 2011).*

The major splicing mechanism in eukaryotes involves a sequence of coordinated steps, each associated with distinct spliceosome complexes (Hoskins et al. 2011). It starts with the formation of the E complex, where U1 snRNP binds to the 5' ss, SF1 binds to the BPS, U2AF2 (U2AF65) attaches to the PPT, and U2AF1 (U2AF35) binds to the 3' ss. This ATP-independent step is crucial for the initial spliceosome assembly. Following this, the ATP-dependent A complex

is formed when SF1 is replaced by U2 snRNP, which pairs with the BPS and is further stabilized by SF3a, SF3b, U2AF2, and U2AF1. The subsequent formation of the B complex involves the recruitment of U5, U4, and U6 snRNPs. U4 and U6 snRNPs pair strongly through complementary RNA sequences, while U5 snRNP associates via protein-protein interactions. Through several conformational changes, the B complex transitions into the pre-catalytic (active) B* complex. During this transition, U1 snRNP dissociates, U6 snRNP binds to the 5' ss, and U4 snRNP is released, allowing U6 and U2 snRNPs to pair. Finally, the C complex facilitates the removal of introns and the joining of exons through two transesterification reactions, completing the splicing process (**Figure 7**) (Will and Lührmann 2011). Interestingly, domain V (DV) of group II introns plays a critical role in catalysis and it shows strong structural and functional similarities to U6 snRNA in the spliceosome. Both DV of Group II introns and U6 snRNA facilitate the coordination of essential metal ions that drive the splicing reaction (Toor et al. 2008; Keating et al. 2010). Furthermore, U2 snRNA, which pairs with the branch-point sequence of the intron, mirrors the function of the branch-point-binding domain (DVI) in Group II introns (Villa et al. 2002; Lambowitz and Belfort 2015). These parallels in both catalysis and structural configuration have led to the hypothesis that group II introns are evolutionary precursors of the spliceosomal machinery, with U6 and U2 snRNAs representing molecular fossils of this group II self-splicing mechanism (Fica et al. 2013; Lambowitz and Belfort 2015).

The minor splicing pathway (U12 splicing pathway), while similar in its fundamental mechanisms, involves different snRNPs and recognizes distinct splice site sequences. Instead of the U1, U2, U4, and U6 snRNPs used in the major pathway, the minor pathway employs U11, U12, U4atac, and U6atac snRNPs. The minor spliceosome targets a unique set of introns, often referred to as U12-type introns, which have different consensus sequences compared to the major U2-type introns. This specialization allows the minor spliceosome to recognize AT-AC splice sites (**Figure 7**) (Patel and Steitz 2003; Turunen et al. 2013). Further research revealed that some introns with GT-AG boundary sequences were spliced by the U12-type spliceosome, and conversely, some introns with AT-AC boundary sequences were spliced by the U2-type spliceosome (Dietrich et al. 1997). Interestingly, U12 type introns lack a PPT and consequently do not require U2AF for intron recognition.

The U12 splicing pathway is believed to be as ancient as the U2 pathway, with origins dating back over a billion years (Wu et al. 1996). Initially identified in animals, U12 introns were later found in plants, protists, and fungi, indicating that the minor spliceosome predates the divergence of these kingdoms (Jackson 1991; Hall and Padgett 1994). This is supported by the discovery of minor introns and homologous components of the minor splicing pathway, such as U11, U12 small nuclear RNAs (snRNAs), and specific minor spliceosome proteins, in protists and fungi (Russell et al. 2006). The presence of these components in protists suggests that the minor spliceosome may be as ancient as the major spliceosome. Several evolutionary models have been proposed which includes the parasitic invasion model (Sharp 1991; Burge et al. 1998; Lynch and Richardson 2002), the co-divergence model (Tarn and Steitz 1997), or the fission/fusion model (Burge et al. 1998). The parasitic invasion model suggests that after the endosymbiotic event in which an archaeal host engulfed bacteria, descendants of two distinct group II introns invaded nuclear genes (Sharp 1991; Burge et al. 1998; Lynch and Richardson 2002; Rogozin et al. 2012). Repeated lysis of the invading bacterium (which developed into mitochondria) could have seeded these introns (Martin and Koonin 2006a; Rogozin et al. 2012), resulting in the fragmentation of group II introns into new U snRNAs. These novel snRNAs could have co-opted proteins from a pre-existing spliceosome (likely the U2-dependent spliceosome) to splice these newly fragmented introns (Burge et al. 1998; Schneider et al. 2002; Basu et al. 2008). The protosplice site analysis of ancient introns in humans and *Arabidopsis thaliana* suggests that the earliest spliceosomal introns were likely U2-type (Basu et al. 2008), leading to the hypothesis that U2 introns were the first to integrate into nuclear genes, followed by U12 introns (Lynch and Richardson 2002; Basu et al. 2008). The discovery of "twintrons", where U12 and U2 introns are sequentially spliced in the same gene complicates the idea that U12 introns evolved after U2 introns. For instance, in the *Drosophila prospero* gene, a twintron splices alternatively between U12- and U2-dependent pathways, producing distinct mRNA isoforms that encode different proteins essential for nervous system development (Scamborova et al. 2004). The regulated switching between U12- and U2-dependent splicing during development highlights the complexity of spliceosomal evolution. The presence of twintrons suggests that U12 introns may not have evolved after U2 introns, contrary to the parasitic invasion model (Tarn and Steitz 1997).

Phylogenetic analyses of minor and major spliceosomal snRNAs and proteins indicate that both splicing systems were present in the last eukaryotic common ancestor (LECA) (Rogozin et al. 2012). U12 introns are found across a wide range of taxa, including plants, vertebrates, insects, Cnidarians, protists, and fungi (Burge et al. 1998; Russell et al. 2006). However, certain groups, such as some fungi (e.g., *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*), nematodes (e.g., *Caenorhabditis elegans*), and some protists, have lost the minor spliceosomal components (Burge et al. 1998). For instance, while some fungi like Ascomycetes and Basidiomycetes lack minor spliceosomal components, other fungi like *Rhizopus* retain both U12 introns and minor spliceosome proteins. This suggests that some eukaryotic lineages lost the minor spliceosomal system during evolution (Russell et al. 2006). The "class conversion hypothesis" proposes that the loss of U12 introns occurred through unidirectional conversion of U12 introns into U2 introns, particularly in the GT-AG subtype of U12 introns (Russell et al. 2006; Moyer et al. 2020). However, this hypothesis does not account for all U12 intron losses, as recent studies show that U12 intron loss is more common than conversion, especially in vertebrates (Lin et al. 2010). Some U12 introns may have been replaced by U2 introns through splice site mutations. For example, in humans, intron 1 of the PTEN gene is a U12 intron, whereas in *C. elegans*, it is a U2 intron. This process has contributed to the reduced prevalence of U12 introns in modern eukaryotic genomes compared to their ancestral counterparts (Burge et al. 1998; Russell et al. 2006; Lin et al. 2010; Moyer et al. 2020).

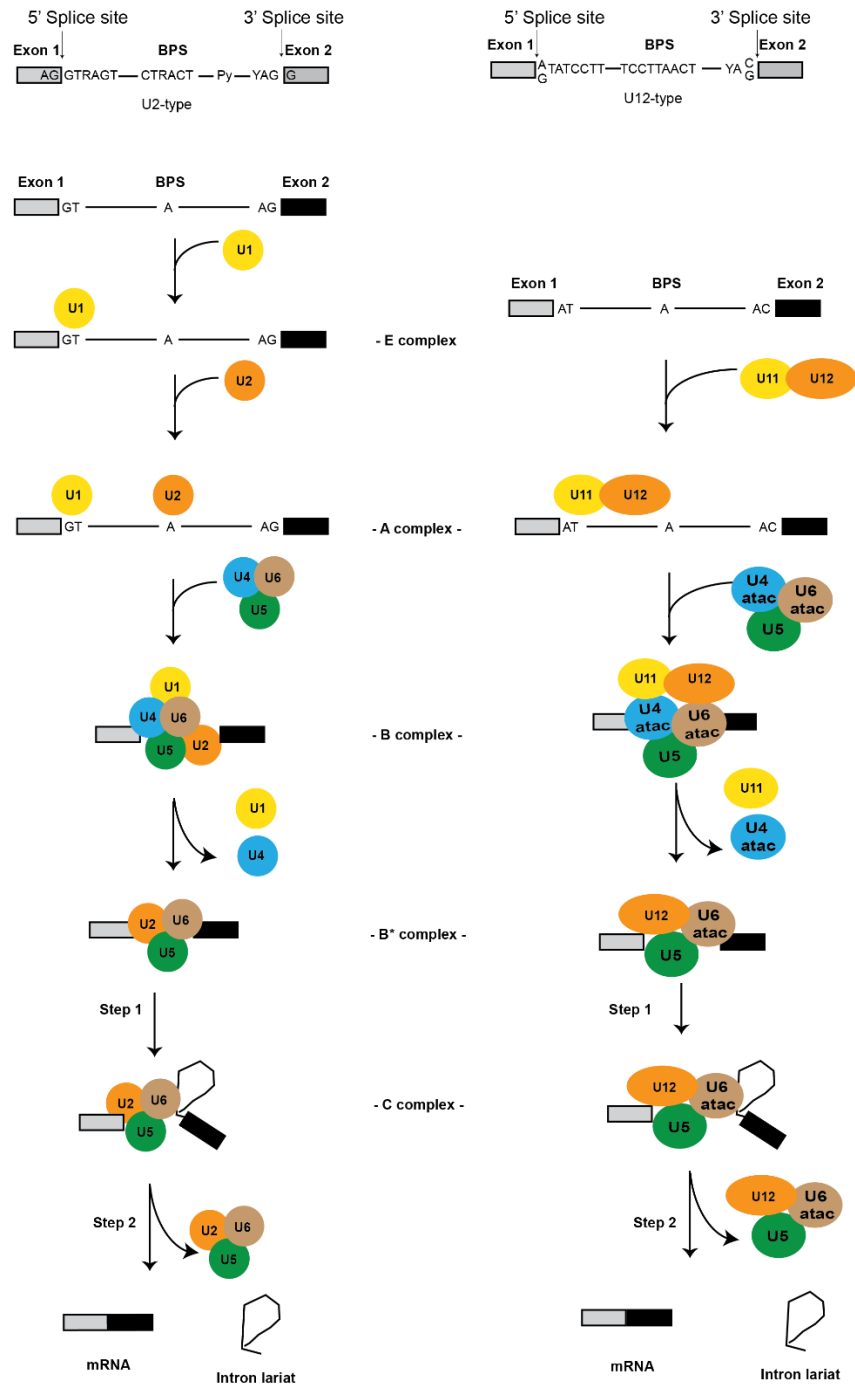


Figure 7. Stepwise assembly of the major (U2-dependent) and minor (U12-dependent) spliceosomes. The interactions of the spliceosomal snRNPs and some selected non-snRNP protein complexes at various stages of spliceosome assembly (complexes E, A, B*, and C) are depicted schematically for both the U2- and U12-dependent spliceosomes. Splice sites are depicted at DNA level. *Created by author inspired from (Will and Luhrmann 2005).*

3.2.1.4.1. Regulation of Splicing

The effectiveness and precision of RNA splicing are heavily dependent on the accurate recognition of the 5' and 3' splice sites (ss). This precision is influenced by several factors, including consensus sequences at exon-intron boundaries and the presence of regulatory elements such as splicing enhancers and silencers. Exonic splicing enhancers (ESEs) and intronic splicing enhancers (ISEs) are short sequences that promote splice site recognition, whereas exonic splicing silencers (ESSs) and intronic splicing silencers (ISSs) are sequences that reduce spliceosome recognition of splice sites, allowing for finely tuned splicing decisions. These regulatory sequences are bound by various proteins, including SR proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs), impacting spliceosome assembly and splicing efficiency (**Figure 8**).

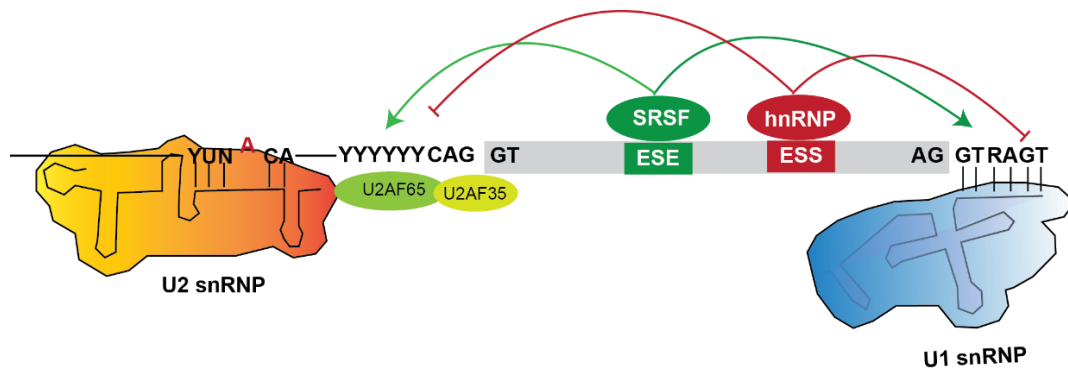


Figure 8. Regulation of splicing by RNA-binding proteins. SR (serine and arginine-rich) proteins [serine and arginine-rich splicing factor (SRSF)]-bound exonic splicing enhancer (ESE) can stimulate splicing at up- and downstream splice sites by facilitating the interaction of U2 auxiliary factor (U2AF) or the U1 snRNP with the RNA. Exonic splicing silencers (ESS) are bound by heterogeneous nuclear ribonucleoprotein (hnRNP) proteins, which antagonize SR proteins and suppress splice sites located up- and downstream. *Created by author inspired by from (Gehring and Roignant 2021).*

3.2.1.4.2. Constitutive and alternative splicing in mRNA diversity

Splicing of multi-exon pre-mRNA transcripts can produce multiple distinct mRNA products (Black 2003; Nilsen and Graveley 2010). There are two primary types of splicing based on event complexity: constitutive splicing (CS) and alternative splicing (AS) (van den Hoogenhof et al. 2016). CS involves the consistent removal of introns from pre-mRNA and the joining of exons to

form a mature mRNA. Whereas, AS allows for the inclusion or exclusion of exons in various combinations, producing a diverse array of mRNA transcripts from a single pre-mRNA (Nilsen and Graveley 2010). Studies have indicated that splicing signals at AS sites are weaker, AS exons are shorter than CS exons, skipped exons more frequently preserve the reading frame in AS than in CS, and orthologous AS events are more conserved than orthologous CS events (Zheng et al. 2005)

3.2.1.4.3. Type of alternative splicing

AS is a mechanism by which a single gene can produce multiple protein products with different functions (Brett et al. 2002; Roy et al. 2013). It was discovered when different mRNA species were found to be encoded by the same gene (Early et al. 1980). High-throughput transcriptome analyses suggest that about 95% of mammalian multi-exonic genes undergo AS (Pan et al. 2008).

AS events can be classified into different splicing pattern, including cassette exon, mutually exclusive exons, retained intron, alternative 5' splice sites, alternative 3' splice sites, alternative promoters, and alternative poly-A sites (Keren et al. 2010). Among them, the most common type of alternative splicing is including or skipping a cassette exon in the mature mRNA. In the cassette exon, internal exons can either be included or excluded from the mature mRNA. In case of intron retention, the excision of an intron can be suppressed, which results in the retention of the entire intron and exons can be extended or shortened through the use of alternative 5' or 3' splice sites. Moreover, alternative promoters and alternative poly-A sites are alternative selection of transcription start sites or poly-A sites and are not due to alternative splicing (**Figure 9**). Among these alternative splicing events, intron retention is generally the most difficult type to detect because it is hard to distinguish from experimental artifacts, because, incompletely spliced transcripts consist of intron fragments and can be inaccurately identified as intron retention. Many genes have multiple alternative splicing events with complex combinations of exons and produce diverse transcript isoforms. For instance, the *Drosophila* Dscam gene contains 20 constitutive and 95 cassette exons. These 95 cassette exons can express 38,016 different mature mRNAs (Missler and Sudhof 1998; Graveley et al. 2004).

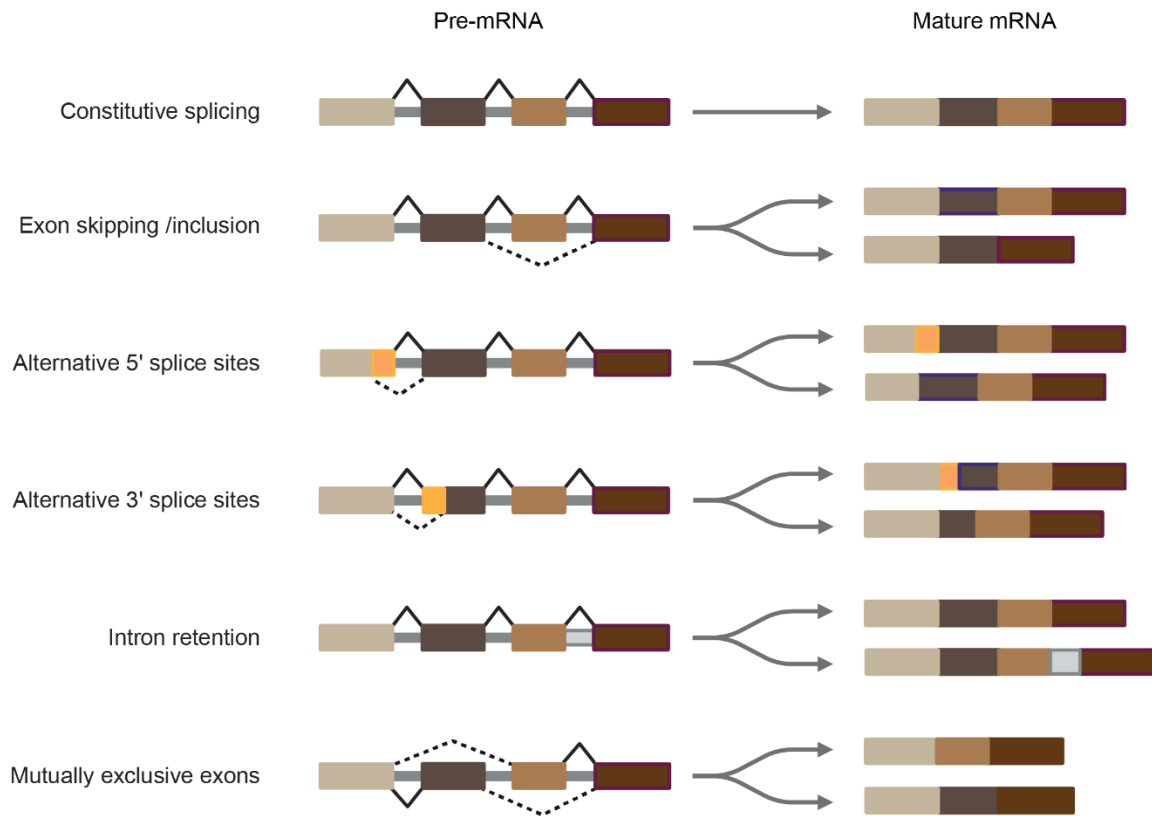


Figure 9. Overview of different types of alternative splicing events.

3.2.3. Intron origin and evolution

The evolution of exon-intron structures in eukaryotic genes and the role of introns have sparked significant debate over the years. This debate has primarily revolved around two main hypotheses: "introns-early" and "introns-late" (Gilbert 1987; Doolittle 1991; Gilbert and Glynias 1993; Logsdon 1998; Lynch and Richardson 2002). The original "strong" introns-early hypothesis suggested that eukaryotic genes inherited almost all their introns from prokaryotic ancestors. According to this view, differences in gene structure among homologous eukaryotic genes were mainly due to differential intron loss (Gilbert 1987). In this scenario, modern prokaryotes lost all their primordial introns and the spliceosome during a process of genome streamlining.

Later versions of the introns-early hypothesis took a more balanced approach, allowing for the emergence of some new introns in addition to the ancient ones (Gilbert and Glynias 1993). On the other hand, the introns-late hypothesis proposed that introns were a novel feature of eukaryotes,

continuously emerging throughout eukaryotic evolution. This perspective suggests that bacteria and archaea have never contained introns or a spliceosome (Stoltzfus et al. 1994; Stoltzfus et al. 1997; Logsdon 1998). Over time, these conflicting perspectives have merged into a synthetic concept suggesting that many introns were present early in eukaryotic evolution (Rogozin et al. 2003; Koonin 2006). Additionally, the 'introns-first' scenario has sought to revive the introns-early idea. This hypothesis proposes that the exons of protein-coding genes emerged from primordial introns interspersed between functional RNA sequences in the RNA world, predating proteins (Poole et al. 1998; Penny et al. 2009). These ongoing debates and evolving theories highlight the complexity and dynamic nature of genomic evolution, underscoring the intricate history of introns in shaping modern eukaryotic genes.

3.3. LncRNAs

LncRNAs are RNA molecules longer than 200 nucleotides that do not encode proteins. Transcribed by RNA polymerase II, lncRNAs share several structural features with mRNAs, including a 5' cap, splicing events, and a polyadenylated 3' end, with about 60% containing poly(A) tails (Sun et al. 2018). Unlike protein-coding mRNAs, lncRNAs are generally expressed at lower levels and exhibit tissue- or cell-type-specific expression patterns (Ma et al. 2013; Waseem et al. 2021; Mattick et al. 2023). They are often multi-exonic and exhibit a high degree of AS, a characteristic that became evident with the advent of high-depth sequencing (Deveson et al. 2018). These RNAs frequently contain a higher proportion of GC–AG splice sites (Abou Alezz et al. 2020), leading to less efficient splicing compared to protein-coding transcripts (Tilgner et al. 2012a; Mele et al. 2017). These features are commonly associated with AS (Garg and Green 2007). Consequently, AS has been shown to significantly impact the function of lncRNAs (Li et al. 2015; Guo et al. 2020; Khan et al. 2021).

3.3.1. Origin and classification of lncRNAs

Classification of lncRNAs can be based on their genomic location relative to protein-coding genes, and they fall into five categories (Latgé et al. 2018) : 1.) Bidirectional lncRNAs have transcription start sites close to a PCG but are transcribed in the opposite direction, typically within 1 kb of the gene's promoter, often influencing in cis the expression of neighboring genes (Wei et

al. 2011). 2.) Intergenic lncRNAs (lincRNAs) are found between protein-coding genes and differ from these genes in terms of their length, number of exons, and how many transcripts are produced (Ulitsky and Bartel 2013). 3.) Intronic lncRNAs are produced from the introns of PCGs 4.) Sense lncRNAs are transcribed in the same direction as the protein-coding genes they overlap with, 5.) while antisense lncRNAs are transcribed in the opposite direction, sometimes partially or fully overlapping with one or more exons (**Figure 10**) (Zhao et al. 2023).

The origins of lncRNAs are diverse and can be attributed to several genetic mechanisms. Mutations in protein-coding genes, such as frameshifts, can disrupt the open reading frames while preserving the expression of their RNA transcripts, effectively converting a PCG into a noncoding RNA gene. Chromosomal rearrangements can create novel expressed noncoding sequences by joining two separate sequences, leading to the formation of new lncRNAs. Additionally, duplications of noncoding RNA sequences can result in repeated elements, increasing the length of the RNA transcript and potentially giving rise to new lncRNAs. Another pathway involves the insertion of transposable elements containing transcriptional start sites into the genome, creating new functional, noncoding RNA sequences (Ponting et al. 2009; Mattick et al. 2023).

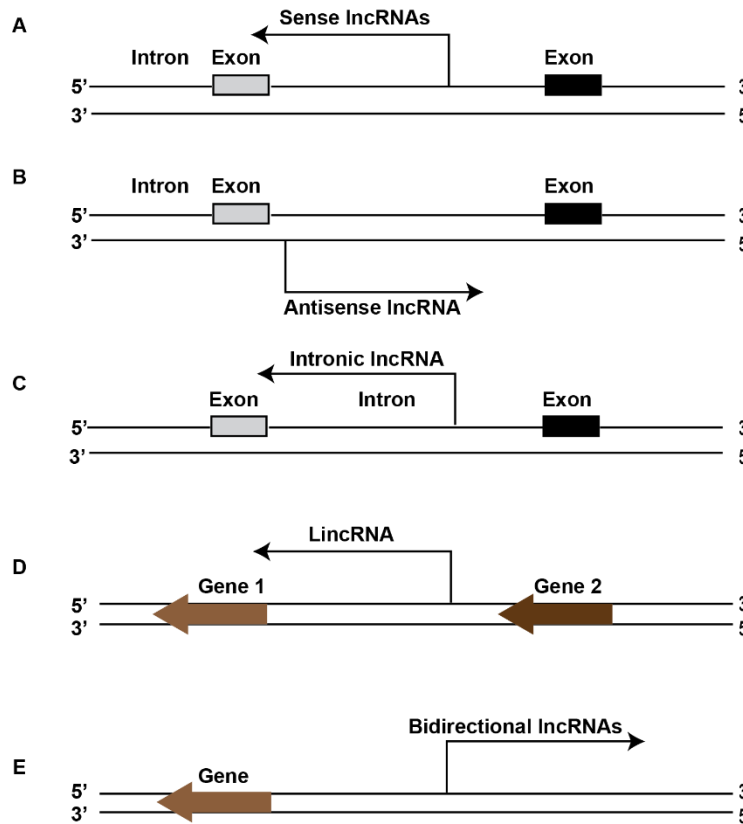


Figure 10. Classification of lncRNAs based on genomic location. A Sense lncRNA: transcribed from sense strand of protein coding gene; B antisense lncRNA: transcribed from antisense chain of protein coding gene; C Intronic lncRNAs: transcribed entirely from introns of protein coding genes; D LincRNAs: entirely derived from interval between two protein coding genes. E Bidirectional lncRNAs: transcribed from the same region opposite to the protein coding genes. *Adopted from (Zhao et al. 2023).*

3.3.2. Localization and Function

LncRNAs were initially discovered for their role in regulating chromatin, and early studies painted a picture of these molecules as mainly nuclear (Clemson et al. 1996; Khalil et al. 2009; Mondal et al. 2010; Tsai et al. 2010; Bridges et al. 2021). For instance, a microarray screen identified XIST, NEAT1, and MALAT1 as some of the most abundant nuclear lncRNAs (Hutchinson et al. 2007). While it's true that more lncRNAs are found in the nucleus than in the cytoplasm (Cabili et al. 2015; Kaewsapsak et al. 2017; Fazal et al. 2019), other studies have shown that there are actually more cytoplasmic lncRNAs than we once thought (van Heesch et al. 2014; Carlevaro-Fita et al. 2016; Bouvrette et al. 2018). Interestingly, while nuclear lncRNAs are generally more abundant,

they tend to be less stable compared to their cytoplasmic counterparts. This instability is believed to reflect their dynamic role in gene regulation, allowing cells to fine-tune gene expression in response to different stimuli, much like transcription factors (Clark et al. 2012; Zuckerman and Ulitsky 2019). Part of this regulation is driven by PABPN1, which helps degrade nuclear lncRNAs by promoting hyperadenylation (Bresson et al. 2015).

lncRNAs perform different functions depending on whether they are in the nucleus or the cytoplasm. In the nucleus, they can regulate transcription by interacting with and remodeling chromatin (Saxena and Carninci 2011; Kugel and Goodrich 2012; Melé and Rinn 2016) and help organize the nuclear architecture by acting as scaffolds (Clemson et al. 2009). In the cytoplasm, their roles shift toward influencing signal transduction, translation, and post-transcriptional gene regulation. For example, cytoplasmic lncRNAs can sequester miRNAs and proteins, affecting their activity and levels (Cesana et al. 2011; Du et al. 2016; Lee et al. 2016; Grelet et al. 2017). They can also modulate protein post-translational modifications (Lin et al. 2016) and play a role in controlling mRNA translation and stability (Gong and Maquat 2011; Carrieri et al. 2012; Yoon et al. 2013; Yuan et al. 2017). An interesting example of this dual functionality is the lncRNA PYCARD-AS1, which operates differently depending on its location in the cell. In the nucleus, it promotes DNA methylation and histone modification at the PYCARD promoter by recruiting specific enzymes, DNMT1 and G9a. But in the cytoplasm, PYCARD-AS1 binds to PYCARD mRNA, preventing ribosome assembly and halting translation (Miao et al. 2019). This shows how the same lncRNA can perform distinct functions depending on the subcellular microenvironment by enabling interactions with different functional protein partners and targets of action.

3.4. Diplonemids: an overview within euglenozoa

Diplonemids are a group within the Euglenozoa, which also includes euglenids, kinetoplastids (like bodonids and trypanosomatids), and symbiontids. This group also includes important pathogens like *Trypanosoma*, *Leishmania*, and *Phytomonas*, as well as photosynthetic organisms like *Euglena*. Unlike these more common relatives, diplonemids were, until recently, rarely found in marine or freshwater environments, with only a few species described. They are generally considered predators of other eukaryotic cells, though some species may be parasitic or symbiotic. The best-known species, *Diplonema papillatum*, (renamed to *Paradiplonema papillatum*) is a sac-

shaped cell with two short flagella and is available from the American Type Culture Collection (Lukes et al. 2015).

These protists share features such as discoidal mitochondrial cristae, a paraflagellar rod, and a similar ancestral feeding apparatus (Triemer and Farmer 1991; Simpson 1997). Additionally, their nuclear DNA contains the unique base 'J' (β -D-glucosyl-hydroxymethyluracil) (Dooijes et al. 2000), and their nuclear transcripts undergo spliced-leader (SL) trans-splicing, where the 5'-end of pre-mRNA is replaced with an SL sequence (Bonen 1993; Liang et al. 2003; McWatters and Russell 2017). These traits were likely inherited from their common euglenozoan ancestor. Current knowledge of the common and unique features related to gene expression and genome architecture among euglenozoans are summarized in **Figure 11** (Valach et al. 2023b).

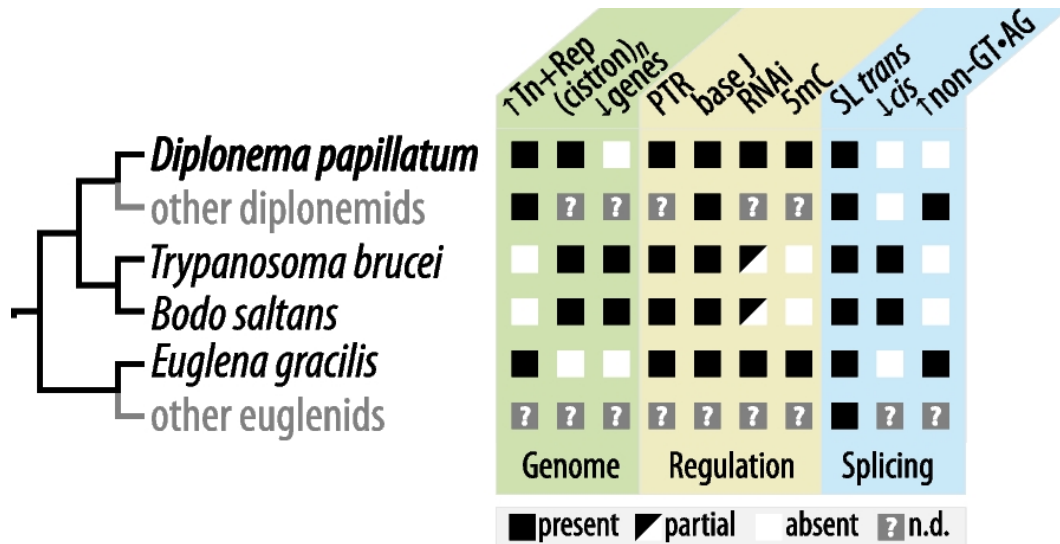


Figure 11: Comparison of genome and gene expression features across euglenozoa. ↑Tn + Rep, conspicuous abundance of transposons and repetitive sequences; (cistron)_n, polycistronic transcription; ↓genes, streamlined gene repertoire; PTR, posttranscriptional regulation of gene expression; base J, base J present in nuclear DNA; RNAi, RNA interference pathway; 5mC, 5-methyldeoxycytosine pathway; SL trans, spliced-leader trans-splicing; ↓cis, few cases of cis-splicing; ↑non-GT•AG, conspicuous abundance of unconventional cis-introns. n.d., not determined. Adopted from (Valach et al. 2023b).

3.4.1. Genome size variation in kinetoplastids and euglenids:

There is a significant difference in genome sizes between kinetoplastids and euglenids. Kinetoplastid genomes are relatively small, around 10^7 base pairs (Berriman et al. 2005; Ivens et al. 2005; Jackson et al. 2016). In contrast, euglenid genomes are much larger, ranging from approximately 3 to 24 Gb (based on DNA content), which is equal to or even greater than the human genome's 3 Gb (Ebenezer et al. 2017). For example, the genome of the diplonemid *P. papillatum* is estimated to be 280Mb (Valach et al. 2023b). Recent estimates of the *E. gracilis* nuclear genome size vary from less than 0.5 Gb to over 2 Gb based on DNA sequencing (Ebenezer et al. 2017; Ebenezer et al. 2019), making it roughly 10 to 100 times larger than the genomes of kinetoplastids like *B. saltans* (40 Mb), *T. brucei* (26 Mb), *L. major* (33 Mb), and *T. cruzi* (28 Mb)(Jackson et al. 2016).

Transcriptome analysis of *E. gracilis* shows approximately 38 Mb of coding sequences (Ebenezer et al. 2017; Ebenezer et al. 2019), compared to kinetoplastid genomes, which have 13-31 Mb of coding sequences (Jackson et al. 2016). This indicates that while the coding sequences of *E. gracilis* are at most three times larger than those of kinetoplastids, the significant difference in genome sizes is likely due to the large amounts of non-coding DNA in euglenids, including introns, repetitive sequences, and mobile genetic elements. The coding regions probably represent less than 1% of the *E. gracilis* genome (Ebenezer et al. 2017; Ebenezer et al. 2019). The size of the ancestral euglenozoan genome remains unknown. However, genome reduction in kinetoplastids and genome expansion in euglenids have likely contributed to the significant difference in genome sizes between these lineages, highlighting their diverse evolutionary paths and adaptations.

3.4.2. Marine Diplonemids: diversity, distribution

Environmental surveys of small subunit ribosomal RNA (SSU rRNA) gene sequences suggest that diplonemids may play a significant role in the deep ocean biosphere (Lopez-Garcia et al. 2001; Lara et al. 2009). Phylogenetic analyses have identified several distinct clades of ‘deep sea pelagic diplonemids’ (DSPDs), which differ from each other and from traditional diplonemids like *D. papillatum*. In an effort to investigate plankton diversity, the Tara Oceans expedition employed a metabarcoding approach, processing samples from 47 locations around the world, focusing on the

photic zone. This effort yielded 580 million DNA sequence reads from the hypervariable V9 region of rDNA (de Vargas et al. 2015). Surprisingly, diplomemid-related sequences were highly abundant in samples from the sunlit ocean, leading to new questions about the ecological roles of diplomemids in marine environments.

Flegontova and colleagues (Flegontova et al. 2016) conducted a comprehensive analysis of diplomemid diversity and distribution, utilizing an expanded dataset from the later stages of the Tara Oceans survey. This study included 850 samples collected from 123 sites, focusing on V9 rDNA metabarcodes from various size-fractionated plankton communities within the mesopelagic zone (200–1,000 m depth). The results were remarkable, revealing that diplomemids accounted for about one-fifth of the eukaryotic operational taxonomic units (OTUs) in the dataset, surpassing metazoans (16%) and dinoflagellates (15%). DNA read counts indicated that diplomemids comprised nearly 15% of eukaryotic cells in the mesopelagic layer, compared to only about 1% in the photic zone. Phylogenetic analysis of all available 18S rRNA sequences longer than 500 bp confirmed that diplomemids form a robust monophyletic group divided into four primary lineages. Approximately 97% of diplomemid sequences were found within the DSPD I clade, suggesting a recent evolutionary radiation due to their high similarity (**Figure 12**). When the short V9 barcodes were mapped onto this reference tree, they all clustered within the established clades, with no new phylogenetic structures emerging. Despite the absence of clear biogeographical patterns, diplomemids appear to be stratified within the water column, underscoring their widespread distribution and significant abundance in marine environments.

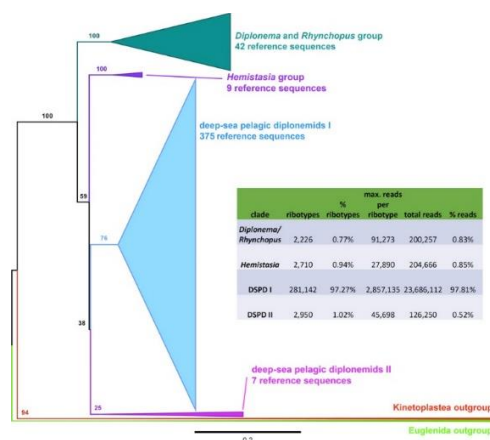


Figure 12. A maximum-likelihood tree based on 433 diplomemid 18S rRNA sequences longer than 500 bp and kinetoplastid and euglenid outgroups. Adopted from (Flegontova et al. 2016).

Gawryluk and colleagues (Gawryluk et al. 2016) explored diplonemid diversity using a complementary approach by applying single-cell genomics to samples from around 100 meters deep in the North Pacific Ocean. They isolated and imaged 92 colorless, flagellated eukaryotes, 10–30 μm in size, using light microscopy and subjected their DNA to multiple displacement amplification. Of the 40 cells identified through SSU rDNA sequencing and phylogenetic analysis, ten were classified as diplonemids (**Figure 13**).

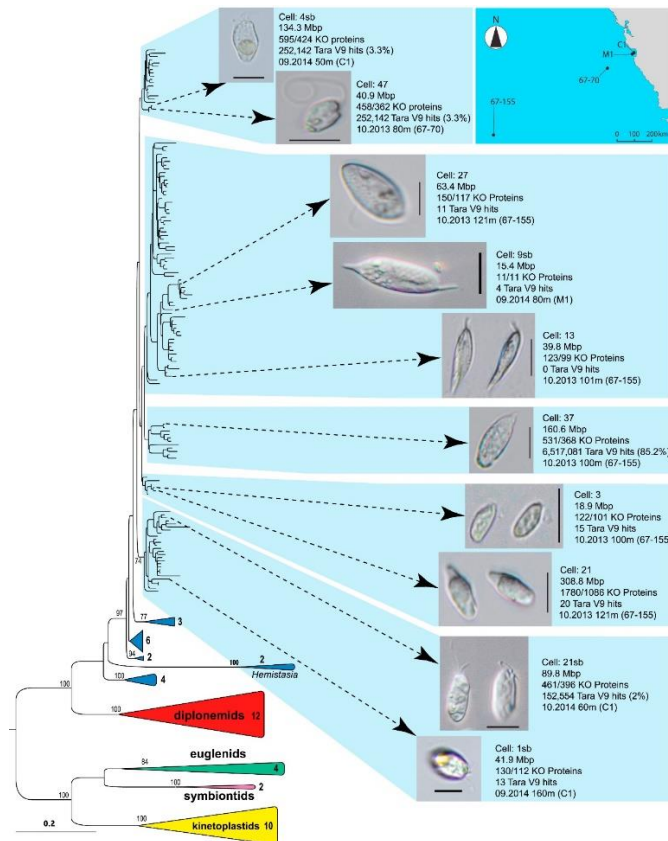


Figure 13. Morphology and evolution of marine diplonemids A euglenozoan-rooted schematic based on maximum likelihood phylogenetic analysis of the diplonemid 18S rRNA gene is shown (1,802 sites; 175 operational taxonomic units [OTUs]). Bootstrap values less than 70 are not shown. Bold numbers refer to the number of taxa present in collapsed clades. Dotted lines connect the phylogenetic placement of each cell's 18S rRNA gene to a photo of that cell. SCG assembly size, number of redundant/non-redundant KO homologs, number of hits to Tara Oceans V9 data (and % of total *Discoba* hits), collection date, station, and depth are presented beside each photo. The scale bars represent 10 μm . “Classical” diplonemids (such as *Diplonema* and *Rhynchopus*) group within the red triangle labeled “diplonemids”; Hemistasia falls within a small group at the base of the planktonic group. A map of collection sites is in the top right corner. *Adopted from* (Gawryluk et al. 2016).

The single-cell genomic data combined with light microscopy observations from Gawryluk et al. provide insights into the ecological roles of these enigmatic protists. Marine diplonemids appear to be flagellated, sexually reproducing predators capable of consuming both prokaryotes (including cyanobacteria) and eukaryotes. Some may also be parasites or symbionts, as indicated by their molecular signatures in the larger plankton size fractions sampled during the Tara Oceans expedition (Lima-Mendez et al. 2015; Lukes et al. 2015; Flegontova et al. 2016). These findings align with the known lifestyle of *Hemistasia*, a recently cultured marine diplonemid that grazes on algae, such as haptophytes and diatoms, as well as small crustaceans (Yabuki and Tame 2015). Phylogenetic analyses show that *Hemistasia* is the closest known relative of the novel marine diplonemids studied by Gawryluk and colleagues for which a laboratory culture has been established (Gawryluk et al. 2016) .

3.4.3. Diversity of cis-spliced introns in euglenozoan nuclear genomes

Despite the limited number of sequenced and annotated euglenid genes, three types of cis-spliced introns have been identified, showcasing the diversity and complexity of these genetic elements (McWatters and Russell 2017).

3.4.3.1. Canonical spliceosomal introns in euglenids

These introns, characterized by their 5'-GT/AG-3' borders, have been found in several euglenid genes, including tubulin (Ebel et al. 1999; Canaday et al. 2001; Milanowski et al. 2014; Milanowski et al. 2016; Ebenezer et al. 2017) and fibrillarin genes (Breckenridge et al. 1999; Russell et al. 2005). They can potentially base pair with U1 small nuclear RNA (snRNA), indicating that they are excised in a U1 spliceosome-dependent manner (Breckenridge et al. 1999). In chloroplast protein genes, conventional (canonical) introns are short and share sequence identity, suggesting a role in adding chloroplast targeting presequences to nuclear genes of endosymbiotic origin (Vesteg et al. 2010).

3.4.3.2. Non-canonical introns in euglenids

Found in various euglenid genes, including the Lhcb gene (encoding the light-harvesting chlorophyll a/b-binding protein) (Muchhal and Schwartzbach 1992), RbcS genes (encoding the small subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase – RuBisCo SSU) (Muchhal and Schwartzbach 1994; Tessier et al. 1995), PetJ (encoding cytochrome c6), PsbO (encoding the 30 kDa protein of the oxygen-evolving complex), and PsbW genes (Vesteg et al. 2010). These introns lack the conserved 5'-GT/AG-3' splicing borders and do not complement U1 snRNA. They lack the structural characteristics of group I and II introns and are often flanked by short direct repeats. Non-conventional (non-canonical) introns can form secondary structures that may help bring the 5' and 3' intron ends together (Muchhal and Schwartzbach 1992; Muchhal and Schwartzbach 1994; Henze et al. 1995; Tessier et al. 1995; Canaday et al. 2001; Milanowski et al. 2014; Milanowski et al. 2016; Ebenezer et al. 2017; McWatters and Russell 2017).

3.4.3.3. Intermediate introns in euglenids: Connecting canonical and non-canonical splicing

Euglenid genes also harbor a third category of introns known as intermediate introns (Canaday et al. 2001). These intermediate introns exhibit features reminiscent of both canonical and non-canonical introns. They possess a 5'-GU or AG-3' borders and have secondary structures that facilitate the joining of intron ends. Recent studies have proposed that the presence of a 5'-GU end in intermediate introns may be a result of evolutionary pressure aimed at preserving a purine at the intron end, rather than representing a transitional stage between canonical and non-canonical introns (Milanowski et al. 2014).

In euglenid species, canonical intron positions in TubA and TubB genes (encoding α - and β -tubulin, respectively) are conserved, while non-canonical intron positions are lineage or species-specific (Milanowski et al. 2014). Canonical introns appear to be removed before non-canonical ones (Guminska et al. 2018). The formation of non-canonical introns is an ongoing process, as evidenced by a recently acquired non-canonical intron in the GapC gene of *Euglena agilis* (Milanowski et al. 2016). Non-canonical introns may have evolved from transposons similar to miniature inverted-repeat transposable elements (MITEs) (Milanowski et al. 2014).

3.4.4. Non-canonical introns in marine diplomonads: insights into their structure, function, and evolution

The nuclear gene structure of diplomonads, particularly within marine clades, remains underexplored. Classical diplomonads, like other euglenozoans, utilize spliceosome-dependent trans-splicing for mRNA processing (Sturm et al. 2001). However, the broader implications of this mechanism in marine clades have only recently been examined. In a significant advancement, Gawryluk and colleagues (Gawryluk et al. 2016) conducted deep sequencing of diplomonad genomes, revealing the presence of SL RNA genes across all marine diplomonad species studied. The conservation of at least one SL-RNA per species, identical to the 39-nt *Diplonema papillatum* homolog, suggests strong evolutionary conservation and the potential utility of SL-RNA gene primers in enriching mRNAs from diverse marine diplomonad species for both environmental and single-cell transcriptomic studies.

Despite this progress, our understanding of other aspects of diplomonad gene structure remains limited. This gap is particularly problematic for developing reliable gene models necessary for interpreting genomic and metagenomic data. Spliceosomal introns, a key feature in eukaryotic gene expression, are sparsely documented in classical diplomonads (Qian and Keeling 2001) and are notably rare in kinetoplastids, the closest relatives of diplomonads. However, more distantly related euglenoids exhibit both canonical spliceosomal introns and a limited number of non-canonical introns that deviate from the typical GT-AG splice boundaries (Muchhal and Schwartzbach 1994; Tessier et al. 1995; Breckenridge et al. 1999; Milanowski et al. 2014).

Gawryluk et al. (Gawryluk et al. 2016) conducted a meticulous analysis of highly conserved genes within single-cell genomic (SCG) assemblies to identify introns in the absence of transcriptome data. They reported a high density of non-canonical introns lacking conventional GT-AG splice boundaries (**Figure 14A**). These introns do not exhibit the typical conservation of splice site sequences; instead, they show a slight overrepresentation of G and C nucleotides at both splice sites (**Figure 14B**). The introns frequently feature extensive base-pairing potential between their 5' and 3' splice sites and often contain short direct repeats (3–6 bp) that complicate the precise determination of intron-exon boundaries (**Figure 14A**). This challenge persists even in the presence of orthologous transcriptome data. The non-canonical introns vary significantly in length,

ranging from 38 to 8,039 nucleotides, with an average length of 409 nucleotides. In some cases, very short exons, such as a 6-bp exon in an alpha-tubulin gene, were observed.

The evolutionary origin of these non-canonical introns remains unclear. They are hypothesized to be spliceosome-independent, as their 5' splice sites do not complement U1 small nuclear RNA (snRNA) (Breckenridge et al. 1999). Interestingly, Gawryluk et al. (Gawryluk et al. 2016) identified candidate U1 snRNA genes in six of the ten SCG assemblies. U1 snRNA typically binds the 5' splice site in cis-spliced spliceosomal introns and also associates with the spliced leader in SL-based trans-splicing. The U1 snRNAs in marine diplomonids are predicted to bind 5' splice site motifs akin to those found in the spliceosomal introns of other diplomonids (Qian and Keeling 2001). However, these motifs are absent in non-canonical introns, suggesting that U1 snRNA might be primarily involved in spliced leader addition rather than canonical splicing in these organisms.

A particularly intriguing aspect of the study is the discovery of a large non-canonical intron within an alpha-tubulin gene, which encodes a reverse transcriptase, a protein typically associated with the mobility of self-splicing introns in other systems, such as yeast mitochondria (Kennell et al. 1993). The non-canonical introns identified may represent a novel and potentially active class of mobile genetic elements that are spliced at the RNA level. Although these introns have been part of the euglenozoan lineage for some time, their recent proliferation within marine diplomonid genomes raises interesting questions about their evolutionary significance. If these introns are indeed mobile elements, they may provide valuable insights into the origin and spread of spliceosomal introns. Nonetheless, their presence poses substantial challenges for future genomic analyses of marine diplomonids, particularly in generating accurate gene models and understanding the full complexity of their genetic architecture.

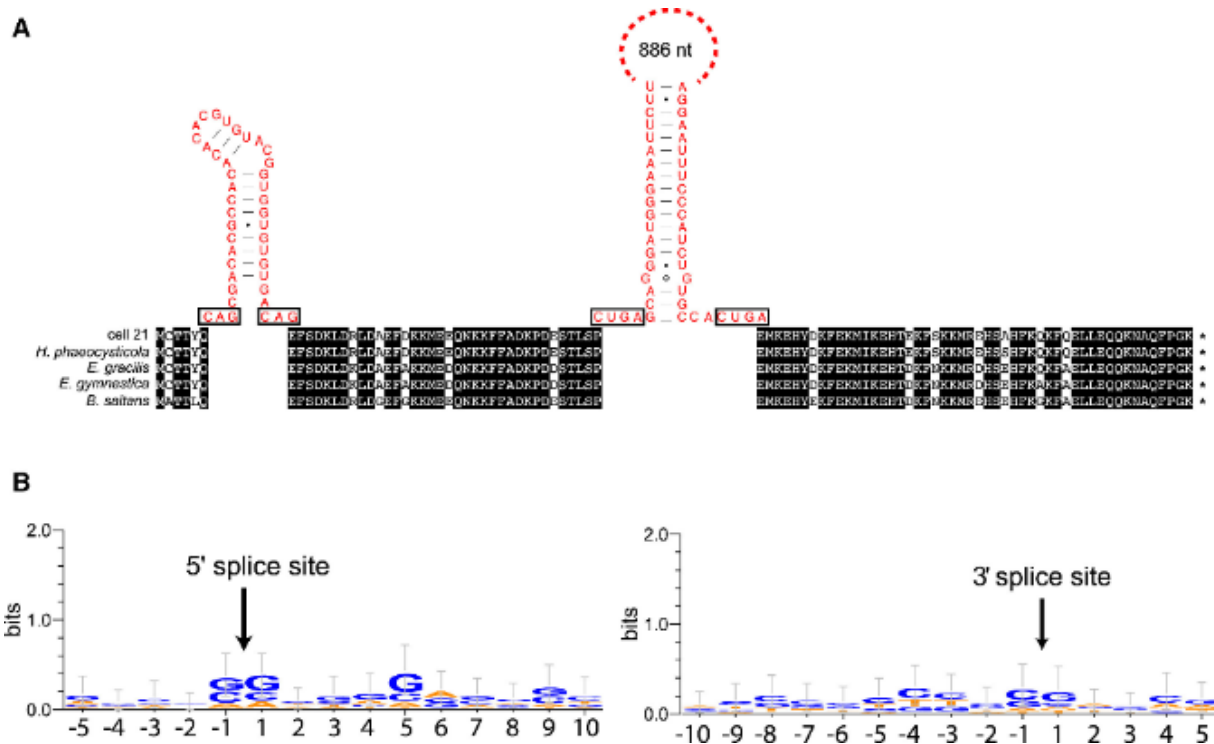


Figure 14. Marine diplomonid genes contain non-canonical introns. (A) Multiple alignment of kinetoplastid membrane protein-11 (KMP-11) showing the position of two non-canonical diplomonid introns in cell 21 (the complete sequence of the 5' intron is shown, but 886 nucleotides from the 3' intron have been omitted for brevity). Extensive potential base pairing interactions between the 5' and 3' intron ends and short direct repeats spanning splice sites (boxed) are shown, which are both common to most observed non-canonical introns. Inferring the exact splice sites is challenging, even in this highly conserved gene, because of the direct repeats: some amino acids are (at least partially) encoded by coding sequence that could be derived from either repeat or both. In the case of the upstream intron, for example, the residues CAG must be in the mature mRNA, but three different pairs of splice sites generate the required sequence. (B) Sequence logos of 15 residues spanning the 5' (left) and 3' (right) splice sites of 44 non-canonical introns (chosen due to particularly obvious capacity to form stable base pairing between 5' and 3' intron splice sites, to aid in their alignment). Splice sites lack canonical GT-AG boundaries but do show a slight elevation of GC. Adopted from (Gawryluk et al. 2016).

4. Project 1 - MATERIAL AND METHODS

Calculation of splice site strength

The human hg19/GRCh37 genome assembly was used as the reference genome with annotations obtained from GENCODE (Release 19) for both, PCGs and lincRNAs (Harrow et al. 2012). Single-exon genes were filtered out. The remaining list with the coordinates of all intron-containing lincRNAs and PCGs was used to extract positions -3 to +6 and -20 to +3 from each 5' and 3'ss using a custom python script. The corresponding sequences were retrieved using BEDTools getfasta (Quinlan and Hall 2010) (v2.17.0), and splice site scores were calculated using the Maximum Entropy Score algorithm (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html) (Yeo and Burge 2004).

Splicing index, exon/intron length and polypyrimidine tract analysis

The splicing index (SI) was computed from RNA-Seq datasets of five different human cell lines (A549, H1-hESC, HepG2, HeLa and MCF7) obtained from the ENCODE project (<https://www.encodeproject.org/experiments/>) (Sloan et al. 2016). Processed and mapped data files (bam files) were downloaded from the ENCODE website: A549 - ENCFF000ELG, H1-hESC - ENCFF000FEH, HepG2 - ENCFF074BOV, HeLa - ENCFF485AHM and MCF7 - ENCFF000HSJ. To establish a list of expressed genes in each cell line separately, we calculated Reads Per Kilobase of transcript per Million mapped reads (RPKM) values for each gene in each of the cell lines. Only genes with RPKMs higher than 0.01 were considered expressed. The top 10% highly expressed PCGs were filtered out to set an upper expression threshold. All lincRNAs with RPKMs above this threshold were removed from further analysis. The SI was calculated as described previously in (Převorovský et al. 2016). Each bam file was divided into two separate files – gapped reads and non-gapped reads. Gapped reads were extracted from the bam files using samtools with an awk command based on CIGAR strings containing the 'N'-label (gap reads) and mapped to the corresponding 5' and 3'ss. Non-gapped reads were extracted based on CIGAR strings without 'N'-label (no gap in read). Alternative exons with overlapping coordinates were excluded from this analysis. To obtain unspliced reads, the last nucleotide in the intron and 25

nucleotides from the exon were extracted at each 3'ss. Reads overlapping the last intronic nucleotide were counted from the extracted non-gapped reads using BEDTools Coverage (v2.17.0)(Quinlan and Hall 2010) with -split and -s options. Custom R scripts were used to determine the SI of all 3'ss for PCGs and lincRNAs. SIs were computed only for 3'ss that contained both, spliced (gapped) and unspliced reads (overlap with the last nucleotide of the intron) and were calculated as the ratio between the number of spliced and unspliced reads by a custom R script.

Gene and exon co-ordinates were obtained from GENCODE v27. BEDTools Subtract (Quinlan and Hall 2010) was used to get intronic co-ordinates by subtracting the exonic co-ordinates from the gene co-ordinate. Custom R script was used to plot the cumulative distribution for exon and intron lengths.

Obtained SI values were ordered and divided into four groups, representing low, medium, high and highest splicing efficiency. From the grouped introns, a region of 40 bp upstream of the 3'ss including the polypyrimidine tract (PPT) was extracted. Sequence logos were produced using WebLogo (<http://weblogo.berkeley.edu/logo.cgi>)(Crooks et al. 2004). Percentages of T, G, C and A nucleotides were calculated for each region and mean percentages computed for each group using custom Perl scripts.

Prediction of secondary structures

To evaluate potential secondary structures of lincRNAs and PCGs, we utilized two alternative approaches. First, we randomly selected 5000 introns from PCG or lincRNAs shorter than 2000 bp. Then we applied RNAfold (<https://www.tbi.univie.ac.at/RNA/>) to predict the base-pairing potential of selected introns using default parameters(Lorenz et al. 2011). To calculate the base pairing probability, we divided the number of base-pairs by the total number of intron nucleotides. Alternatively, we randomly selected 5000 introns from PCG or lincRNAs and calculated minimum free energy of 100 bps around 3'ss and 5'ss using RNAfold(Lorenz et al. 2011).

Srsf-binding and splicing silencer motif prediction

SRSF protein and splicing silencer consensus motif distributions around splice sites were visualized using seqPattern Bioconductor R package (Gentleman et al. 2004). The references for SR binding consensus motifs with IUPAC nucleotide ambiguity codes are listed in the Supplementary data. Windows of 200 bp (-100 to +100) were extracted around the 5'ss and 3'ss

for both, PCGs and lincRNAs. The `import.bed` function was used from the `Rtracklayer` Bioconductor R package (Lawrence et al. 2009), and the corresponding sequences were extracted using the `getseq` function. The `getPatternOccurrenceList` function of `seqpattern` R package was used to obtain the position of each consensus pattern occurrence for each sequence. A total number of motifs was counted for each position for the given region. Binding site densities were calculated for each position separately as the number of motifs divided by the number of analyzed sequences. The visualization was done in R (version 3.5.1) with the `ggplot2` (Wickham 2016) package for the final graphical output.

iCLIP library preparation

HeLa cells stably expressing SR proteins tagged with GFP at endogenous levels from bacterial artificial chromosomes (BACs) were grown to 90% confluency and irradiated once with 150 mJ/cm² UV light (254 nm). Protein G Dynabeads coupled with goat anti-GFP antibodies (D. Drechsel, MPI-CBG, Dresden) were used for immunopurification. Crosslinked, partially digested RNAs to lengths of 60-150 nucleotides were immunopurified using anti-GFP antibodies, reverse transcribed to generate cDNA libraries and subjected to high-throughput sequencing on Illumina HiSeq2000 (single-end 75 nucleotide reads) (Botti et al. 2017). To normalize for transcript levels, total RNA was isolated from HeLa WT cells (2 replicates) and sequenced after rRNA depletion (RiboMinus) on the same Illumina HiSeq2000 machine (single-end 75 nucleotide reads). The raw data were deposited in the GEO database under accession number GSE113814.

Analysis of iCLIP and RNA-seq data

Adapters and barcodes were removed from the iCLIP reads before mapping to the human hg19 genome assembly (Ensembl59 annotation) using Bowtie version 0.10.1. Uniquely mapped reads were used to extract the crosslink site (1st nucleotide of the read) and the statistical significance of binding events ($FDR < 0.05$) calculated and compared to randomized co-transcribed regions (König et al. 2010). To obtain comparable numbers of significant binding sites (CLIP tags) replicates that correlated best were pooled.

RNA-Seq reads from HeLa cells (2 replicates) were mapped to the hg19 reference using TopHat v2.1.0, where the maximum number of multi mappers (max-multihits) was set to 1. Both mapped replicates were merged using `samtools merge` option and RPKMs were calculated for each

transcript using CoverageBed(Quinlan and Hall 2010). Transcripts with RKPM values lower than 0.01 were considered as not expressed and were excluded from the analysis. PCGs were filtered to match the expression range of lincRNAs. For this, all lincRNAs were sorted according to their RKPM values into 20 bins. Subsequently, PCGs were randomly selected to match the number and expression levels of lincRNAs in each bin and were further used for analyses as expression-matched PCGs. For each expressed transcript, a window of 200 bp around each 5'ss and 3'ss (-100 to +100) was selected and intersected with the iCLIP data using intersectBed(Quinlan and Hall 2010). The total number of CLIP-tags was counted for each region. Mean SR protein binding was determined as the sum of CLIP-tags divided by the number of 200 bp regions used for the analysis. Metagene iCLIP profiles of the 200 bp regions were generated using R/Bioconductor packages (Gentleman et al. 2004). First, the input BED file (containing coordinates of all expression-matched transcripts) and the input GFF file (containing genomic reference regions from GENCODE v19) were imported using rtracklayer(Lawrence et al. 2009). The GFF file was further processed using GenomicFeatures (Lawrence et al. 2009) to extract the intron features (using the GenomicFeatures::intronsByTranscript function). 200bp around the beginning and the end of introns were selected. iCLIP data coverage was considered as one for each position. The iCLIP counts were converted into a GRanges object using the GenomicRanges (Lawrence et al. 2013) and overlapped with the 200 bp regions (using GenomicRanges:: findOverlaps function). Plots were generated using the matplot function in R.

To calculate the cumulative distribution of splicing efficiencies of PTBP1/U2AF2 bound lincRNAs and the sequence compositions of PPTs bound and unbound by U2AF2, the 40 bp regions upstream of 3'ss were intersected with iCLIP data of PTBP1 and U2AF2 (Xue et al. 2009; Zarnack et al. 2013) using the intersectBed program. SIs were calculated as described above using RNA-Seq data from Xue et al. (2009) and Zarnack et al. (2013). Cumulative frequency of SIs for PTBP1/U2AF2-bound and unbound transcripts at 40bp upstream of 3'ss were plotted using the R script.

Mapping of eCLIP sequence data

For mapping eCLIP sequencing data from HepG2 cell line (BioRxiv: <https://doi.org/10.1101/179648>) we used GENCODE (GRCh38.p7) genome assembly and the STAR alignment (version 2.4.2a) using the following parameters from ENCODE pipeline: STAR

--runThreadN 8 --runMode alignReads --genomeDir GRCh38 Gencode v25 --genomeLoad LoadAndKeep --readFilesIn read1, read2, --readFilesCommand zcat --outSAMunmapped Within --outFilterMultimapNmax 1 --outFilterMultimapScoreRange 1 --outSAMattributes All --outSAMtype BAM Unsorted --outFilterType BySJout --outFilterScoreMin 10 --alignEndsType EndToEnd --outFileNamePrefix outfile as described in Haberman et al. (2017); and Chakrabarti et al. (2018). All uniquely mapped reads were corrected for over-amplification of PCR duplicates, by using a python script 'barcode collapse pe.py' available on GitHub (<https://github.com/YeoLab/gscripts/releases/tag/1.0>), which is part of the ENCODE eCLIP pipeline (<https://www.encodeproject.org/pipelines/ENCPL357ADL/>). The visualization part was done in R (version 3.4.1) together with the ggplot2 (<https://github.com/hadley/ggplot>) (Wickham 2016) and the smoother (<https://github.com/config-i1/smooth>) package for the final graphical output. Each density graph shows a distribution of raw crosslinking positions of cDNA-starts relative to splice site positions. Gaussian filtering with 10 nt window size was used for the final smoothing of each density line. A set of expression-matched PCGs was created the same way as in the case of iCLIP analysis, but HepG2 RNA-seq data from ENCODE (HepG2 - ENCFF074BOV) were used.

5. Project 2 - MATERIALS AND METHODS

Cell cultivation, RNA and DNA isolation for library preparation and sequencing

The axenic cultures of *S. specki*, *L. lanifica*, *R. humris*, *D. japonicum*, *A. motanka* and *N. karyoxenos* were grown in an artificial medium composed of 3.6% sea salts (Sigma-Aldrich), 1% (v/v) heat-inactivated horse serum (Sigma-Aldrich), 0.025 g/l LB broth powder (Amresco) as described previously (Tashyreva et al. 2018; Prokopchuk et al. 2019). *P. papillatum* (ATCC 50162) was cultivated axenically following a protocol described previously (Kaur et al. 2018).

For each diplomonid species, DNA and RNA for genomic and RNA-seq sequencing were isolated using the standard phenol-chloroform protocol and the Nucleospin RNA isolation kit (Macherey Nagel), respectively. Paired-end 250 bp Illumina MiSeq reads were obtained using the standard TruSeq protocol. For transcriptome sequencing, paired-end 100 bp reads were sequenced on Illumina HiSeq 4000 platform using the standard TruSeq protocol. The sequencing was carried out by Genome Quebec.

***De novo* genome and transcriptome assembly and annotation**

The overlapping paired-end DNA reads were merged using bbmerge (BBtools v.38.90) (Bushnell et al. 2017) with default settings. The combination of merged and not merged reads was used for the genome assembly using SPAdes v.3.15.5 (Prjibelski et al. 2020) under default parameters. Similarly, for paired-end RNA-seq reads, bbmerge was employed, followed by the assembly using rnaSPAdes (Bushmanova et al. 2019). The transcriptome assembly was subsequently aligned to the assembled genome using GMAP (version 2021-02-12) (Wu and Watanabe 2005), utilizing the parameter `--canonical-mode=0` to generate preliminary gene models for each species. Additionally, paired-end RNA-seq reads were aligned to their respective assembled genomes. STAR aligner v.2.7.8a (Dobin et al. 2013) was utilized for this purpose, with and without reference annotation, using a two-pass mode. Other parameters were: `alignIntronMin=12`, `outFilterMismatchNoverReadLmax=0.02`, `scoreGapNoncan=-4`, `scoreGapATAC=-4`, and `alignIntronMax=10,000`. The output of STAR alignment, both with and without gmap reference annotation, was utilized for transcript reconstruction via StringTie v.2.1.4 with `-rf` flag and guided by reference annotation.

Subsequently, both GTF files derived from StringTie, representing transcript annotation, were used as an input for the gffcompare program v.0.12.2 with the strict-match parameter to generate consensus annotations. Transcripts exhibiting "matching" intron chains with the class code '=' were selected for each species and the introns lacking strand information were excluded from the analysis. Overlapping introns were filtered during the analysis.

In addition, we re-annotated *A. montanka* and *N. karyoxenos* using stringent criteria to refine the alignment process, considering the high abundance of non-canonical introns in both species. The alignment was performed with the following STAR parameters: alignIntronMin 20, alignIntronMax 10000, outFilterMismatchNoverReadLmax 0.02, outFilterMismatchNmax 3, outFilterMultimapNmax 1, scoreGapNoncan -6, scoreGapATAC -6, chimSegmentMin 20, chimJunctionOverhangMin 20, outSJfilterOverhangMin 6 6 6 6, outSJfilterDistToOtherSJmin 5 5 10 10. The alignment was performed in "End-to-End" mode (alignEndsType EndToEnd). Similar to previous run the output of the STAR alignment, both with and without gmap reference annotation, was used for transcript reconstruction with StringTie v2.1.4. The following parameters were applied: m 200, f 0.2, j 8, g 50, c 3, A 5 and rf. Both GTF files generated by StringTie, which represent the transcript annotations, were used as input for the gffcompare program with the strict-match setting to create consensus annotations. For each species, transcripts with "matching" intron chains (class code '=') were selected, and introns without strand information were excluded from the analysis. Overlapping introns were filtered during the analysis. Both annotations produced similar results; however, the second annotation run was selected for downstream analysis of *A. montanka* and *N. karyoxenos*.

To further refine the results, split reads in the BAM files representing exon-exon boundaries were counted using BEDTools v.2.30.0 intersect function, and the boundaries supported by fewer than 15 split reads were discarded. The agat_sp_add_introns tools v.0.7.0 (https://agat.readthedocs.io/en/latest/tools/agat_sp_add_introns.html) was employed to augment intron features to the consensus GTF file. Subsequently, 5' splice sites (3 bp upstream and 6 bp downstream) and 3' splice sites (20 bp upstream and 3 bp downstream) were extracted from the intron coordinates as defined previously (Yeo and Burge 2004; Krechnáková et al. 2019). The corresponding sequences were retrieved using BEDTools (Quinlan and Hall 2010) function 'getfasta'. Sequence logos for 5' and 3' splice sites were generated using WebLogo (Crooks et al. 2004).

Furthermore, to verify the accuracy of our intron prediction, we utilized raw transcriptome and genome sequencing data, as well as the genome assembly and annotation of *P. papillatum* (Valach et al. 2023b). First, raw transcriptomic and genomic reads were analyzed using our pipeline to identify introns in *P. papillatum*, as described above for other diplonemids. Next, sequence logos were created using WebLogo (Crooks et al. 2004) as described above for the splice sites obtained using our pipeline and those derived from the previously published genome annotation of *P. papillatum* (Valach et al. 2023b).

snRNA identification and secondary structure prediction

Rfam covariance models were downloaded from Rfam (<https://rfam.xfam.org/>). The cmprss (INFERNAL v.1.1.3) program was run on Rfam.cm to prepare an input for cmscan program. Small nuclear RNAs (snRNAs) were identified using Infernal's cmscan program (Nawrocki and Eddy 2013) with the parameters 'rfam,' 'cut_ga' and 'nohmmonly'. Secondary structure models for snRNAs were generated using the R2R program v.1.0.5 (Weinberg and Breaker 2011).

Identification of core spliceosome proteins

We established a custom snRNP (small nuclear ribonucleoprotein) database by acquiring core spliceosome proteins associated with U1, U2, U4, U5, and U6 snRNAs from the National Center for Biotechnology Information at <https://www.ncbi.nlm.nih.gov/>. To identify potential candidate proteins, we conducted blastx searches (Altschul et al. 1990) using the above mentioned proteins as queries and *de novo* assembled transcriptomes from each species as a database with the following parameters: *e*-value of $1e-10$, subject_besthit, max_target_seqs 3, and qcov_hsp_perc 25. To ensure the robustness of our findings, we executed reciprocal blastp searches (Altschul et al. 1990) with the following criteria: *e*-value of $1e-10$, max_target_seqs 1, and qcov_hsp_perc 30. Finally, we generated multiple sequence alignments of the candidate spliceosomal proteins using ClustalX v.2.1 (Thompson et al. 2002).

Calculation of base-pairing probability

To investigate the base-pairing propensity between the 5' and 3' ends of introns across various species, intron sequences were extracted using agat_sp_extract_sequences program (https://agat.readthedocs.io/en/latest/tools/agat_sp_extract_sequences.html). We extracted 9-base pair substrings from both the 5' and 3' ends of each intron, concatenating them into subsequences of 18 base pairs. Subsequently, we employed the RNAfold v.2.4.18 (Lorenz et al. 2011) to predict the secondary structure of these sequences. Base-Pairing Probability (BPP) was calculated using

the R programming language (Team 2021), where we divided the total number of base-pairing nucleotides by the total number of nucleotide bases for each species. The ggplot2 package was used to create boxplots. The secondary structure prediction of introns were performed using RNAstructure v.6.4 (Reuter and Mathews 2010). For species *S. specki*, *L. lanifica*, *D. japonicum*, and *R. humris*, non-canonical introns (1.56–2.89%) were filtered out due to their extremely low representation to minimize noise and ensure reliable BPP estimates.

Phylogenetic tree construction

We used a computational pipeline to find the evolutionary relationships inherent within canonical and non-canonical splice sites (5'ss and 3'ss) of *A. motanka*, *N. karyoxenos* with other marine diplonemids (*S. specki*, *L. lanifica*, *D. japonicum*, *R. humris*, and *P. papillatum*). Non-canonical introns were excluded for species *S. specki*, *L. lanifica*, *D. japonicum*, and *R. humris*, as they were poorly represented, allowing the phylogenetic analysis to focus on robust evolutionary signals. Initially, sequences were extracted from positions -3 to +9 at the exon-intron junction (consisting of 3 exonic positions and 9 intronic positions) to represent the 5'ss motif. Similarly, positions -20 to +3 were extracted from the intron-exon junction (comprising 20 intronic positions and 3 exonic positions) to represent the 3'ss motif (Yeo and Burge 2004; Krchnáková et al. 2019). The conserved dinucleotide intron termini were removed from both 5'ss and 3'ss to prevent hiding of signals from other splice site positions. Subsequently, Bioconductor package Biostrings v.2.58.0 (Pagès et al. 2024) used for construction of Position Weight Matrices (PWMs) from the extracted canonical and non-canonical sequences at both 5'ss and 3'ss. Next, a distance matrix was computed using the PWMs, capturing the evolutionary distances between sequences. Hierarchical clustering was then applied to the distance matrix using the hclust function from the ape package v.5.7.1. (Paradis and Schliep 2019). Finally, the hierarchical clustering result was converted into a phylogenetic tree using the 'as.phylo' (Paradis and Schliep 2019) function, and the resulting tree was visualized using the 'plot' function. Furthermore, we constructed the species tree using FastTree implemented in OrthoFinder v.2.5.4 (Emms and Kelly 2019). For that, protein sequences from *A. motanka*, *N. karyoxenos*, *S. specki*, *L. lanifica*, *D. japonicum* and *R. humris* were obtained by predicting open reading frames (ORFs) from their assembled transcripts using OrfPredictor (Min et al. 2005). Protein sequences for *P. papillatum* originate from the recently published genome annotation (Valach et al. 2023b). The species tree was generated by integrating OrthoFinder gene trees with

the STAG (Emms and Kelly 2018) and STRIDE (Emms and Kelly 2017) algorithms. Phylogenomic tree was visualized with iTOL v.5 (Letunic and Bork 2021).

6. Results - Project 1

5' ss sequence is important for lincRNA splicing

To investigate whether 5'ss strength is a general determinant of lincRNA splicing, we analyzed available RNA-Seq data from five different human cell lines (embryonic stem cells H1-hESC, lung carcinoma A549, cervix carcinoma HeLa, liver carcinoma HepG2 and breast cancer cell line MCF7). To avoid a potential overlap with PCGs, we focused on lincRNAs only. We first selected all PCGs and lincRNAs expressed in a particular cell line (RPKM>0.01) and filtered out the top 10% of highly expressed PCGs as an upper expression threshold to avoid bias from highly expressed genes. All lincRNAs with RPKM values above this threshold were also removed from further analyses to keep lincRNAs and PCGs within the same expression range for each cell line. Splicing indices (SI) were then determined for each individual intron in each cell line separately (**Figure 15**). In total, we analyzed between 1,054-1,770 lincRNA and >77,000 PCGs introns. Although the lincRNA expression is highly cell-specific, we found that lincRNAs were, in general, less efficiently spliced (lower SI) compared to PCGs in all tested cell lines (**Figure 15**).

To evaluate 5'ss strength in differently spliced genes, we categorized lincRNA and PCG introns into four groups based on their splicing efficiencies (increasing SI), and we calculated the mean 5'ss maxent score (MES) for each group (**Figures 16 and 17**). We found a positive correlation between 5'ss strength and splicing efficiency of lincRNAs in four tested cell lines (Pearson's correlation coefficients 0.67-0.94) while no such correlation was found for PCGs (Pearson's correlation coefficients -0.75-0.57) (**Figure 17**). These results suggest that lincRNA splicing is more dependent on 5'ss strength than splicing of PCGs.

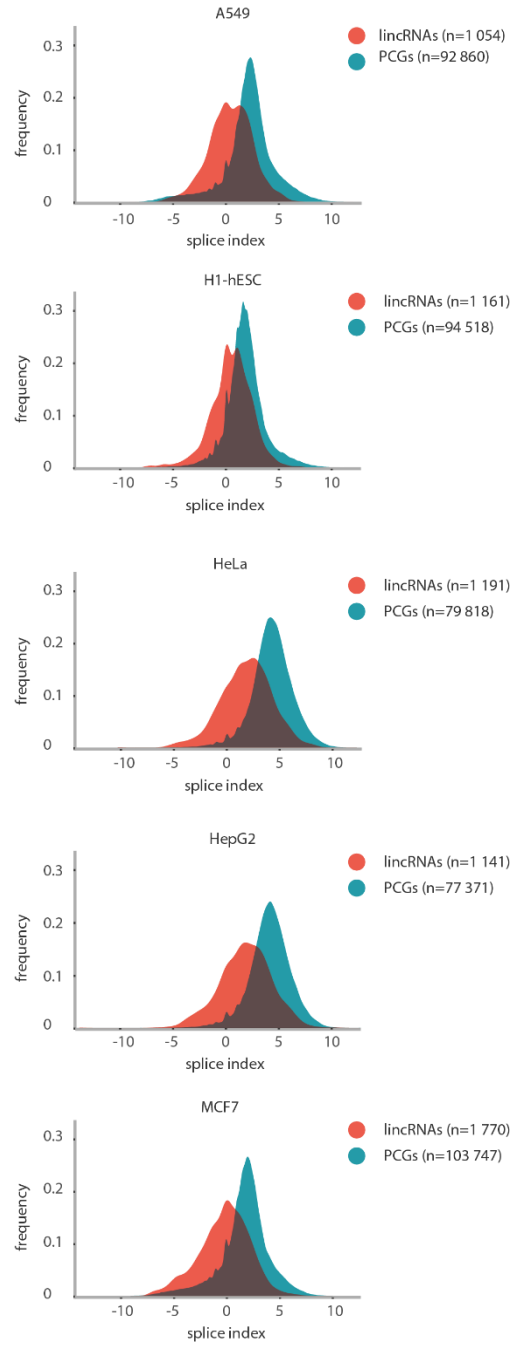


Figure 15. Splice index scores of introns in lincRNAs and PCGs in five different human cell lines; n = the number of introns.

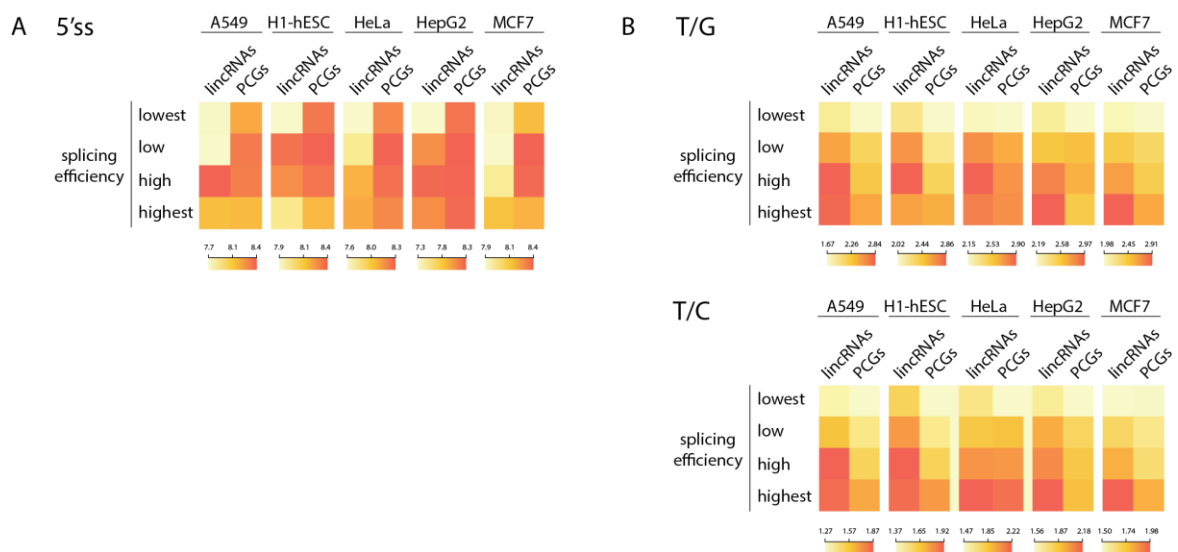


Figure 16. LincRNA splicing efficiency correlates with the strength of 5'ss and PPT. LincRNAs and PCGs were divided into four groups based on their splicing efficiency containing an equal number of transcripts. (A) 5'ss strength or (B) thymidines over guanines (T/G) and thymidines over cytosines (T/C) ratios frequencies in the PPT region was calculated for lincRNAs and PCGs in different cell lines. The color schemes representing the distribution of data are shown under the heat maps. See also Figures 17-20 for further analyses.

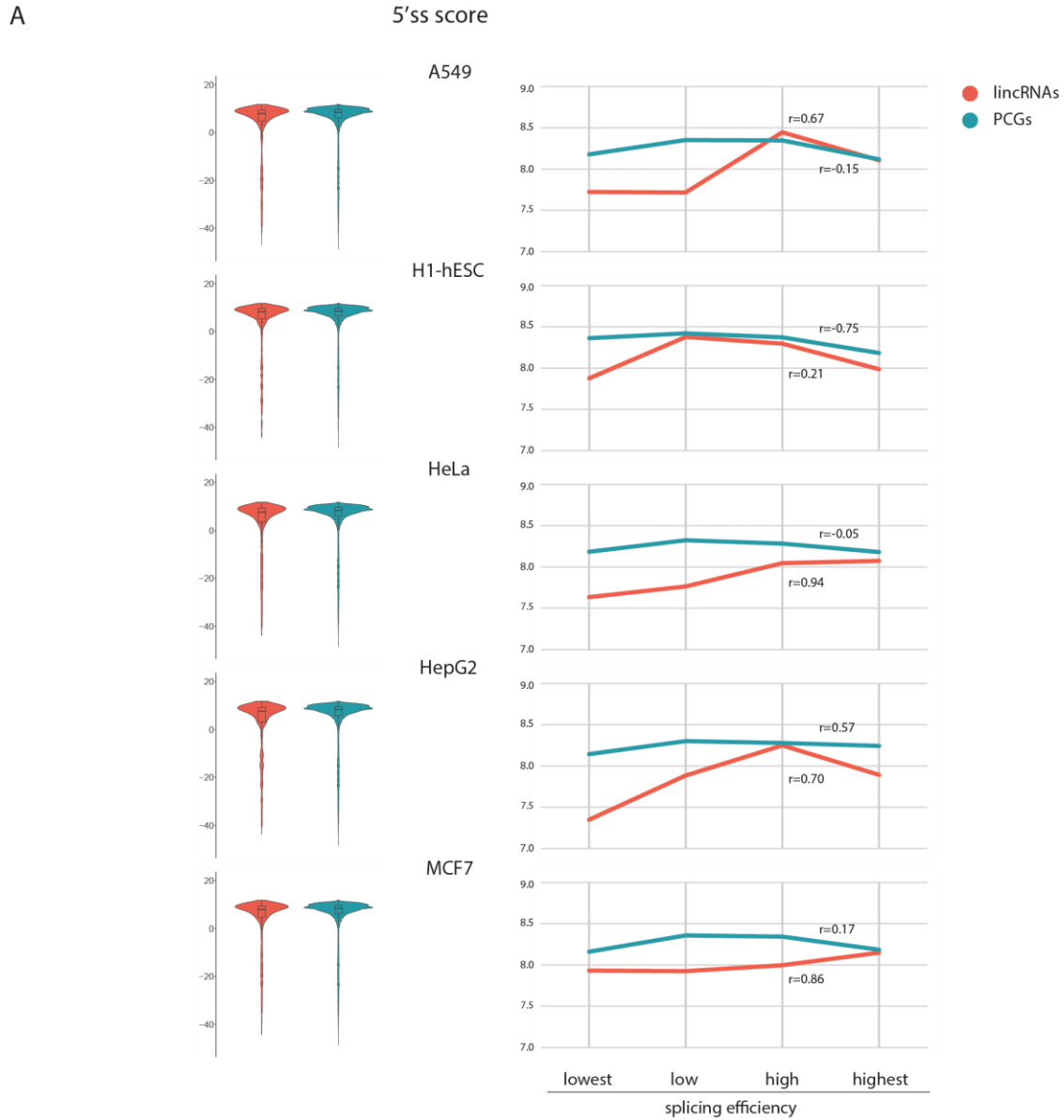


Figure 17. (A) 5'ss score distribution in lincRNAs and PCGs in five different cell lines (left panel); 5'ss scores divided according to splicing efficiency of lincRNAs and PCGs (right panel); r - Pearson's correlation coefficient between 5'ss score and splicing index. (B) A table of mean splicing indexes (SIs) of lincRNAs and PCGs in four groups divided according to the splicing efficiencies.

Polypyrimidine tract sequence determines the splicing efficiency of lincRNAs

In addition to 5'ss, we also compared sequences at the 3' end of introns. A detailed analysis of PPT sequences (nucleotides -40 to -1) revealed slightly better conservation of cytidines/thymidines (C/T) nucleotides at position -3 of the YAG sequence in PCGs in all five tested cell lines (**Figure 18**). Interestingly, we found that the stretch of Ts within the PPT of lincRNAs is longer than in PCGs (**Figure 18**). In line with this finding, a higher number of Ts in lincRNA genes versus PCGs was observed in a recent study analyzing lincRNA splice-site strengths (Mele et al. 2017).

To better understand the role of PPT length and T content, we used the categorized lincRNA and PCG introns based on their splicing efficiencies and calculated the ratios of Ts over guanidines (T/G) or cytidines (T/C) in the PPT region for the five human cell lines (**Figure 16B** and **Figure 19-20**).

The correlation between T content and splicing efficiency was high in both lincRNAs and PCGs (Pearson's correlation coefficients between 0.66 - 0.99; **Figure 19**). However, in all cases, the well-spliced lincRNAs (third and fourth quartiles) contained a higher number of Ts than PCGs (**Figure 16B** and **Figure 19**). While the average SI of best-spliced lincRNAs reached only 60-76% of PCG SI (dependent on the cell line, **Figure 17B**), the T/C and T/G ratios in this group of lincRNAs were higher by 2-13% and 2-16% respectively (**Figure 16B** and **Figure 19**) suggesting that the T content in the PPT can play an important role in lincRNA splicing. Next, we analyzed how splicing factors, which preferentially bind U-rich sequences in the PPT, interact with lincRNAs and how their binding affect lincRNA splicing. We focused on U2AF2 (U2AF65), hnRNP C and PTBP1, which were all shown to bind to the U-rich sequences in the PPT (Mulligan et al. 1992; Wagner and Garcia-Blanco 2001; König et al. 2010; Zarnack et al. 2013). We utilized publicly available eCLIP (enhanced crosslinking and immunoprecipitation) data from HepG2 cells (BioRxiv: <https://doi.org/10.1101/179648>) and iCLIP (individual nucleotide-resolution crosslinking and immunoprecipitation) data from HeLa cells (Xue et al. 2009; Zarnack et al. 2013) and compared splicing efficiencies of lincRNAs associated/not associated with these proteins. hnRNP C iCLIP data revealed the very low association of hnRNP C with lincRNAs (data not shown), and we included only the hnRNP C eCLIP data analysis. In the U2AF2 data set, lincRNAs bound by U2AF2 tend to have a higher T content in PPT, longer PPTs and significantly higher splicing efficiencies than lincRNAs that are not bound by U2AF2 (**Figure 21 A-C**), in agreement with a published analysis (Mele et al. 2017). In contrast, PTBP1-bound and hnRNP C-

bound lincRNAs were spliced as efficiently as unbound lincRNAs (**Figure 21 D-E**). These data suggest that U2AF2 binding improves lincRNA splicing efficiency, while PTBP1 and hnRNP C binding do not. Altogether these results are consistent with the model that inefficient U2AF2 binding is one of the key factors that reduces splicing efficiencies of lincRNAs.

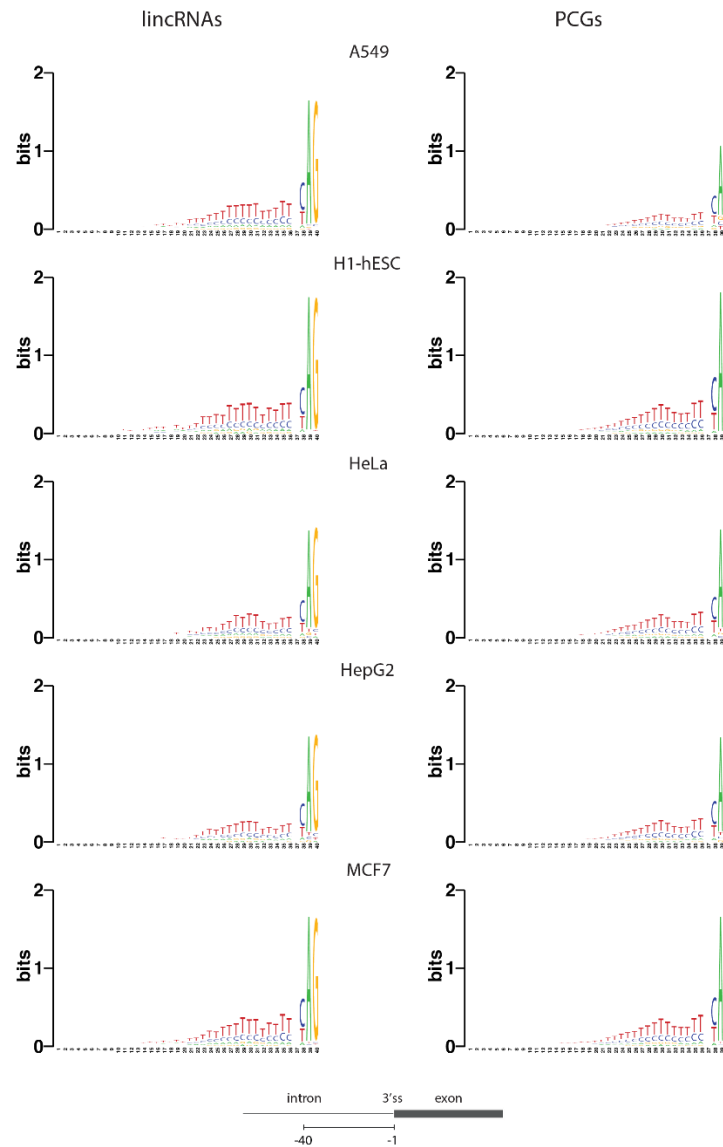


Figure 18. Sequence logos of 3'ss and PPT regions in lincRNAs and PCGs in different cell lines.

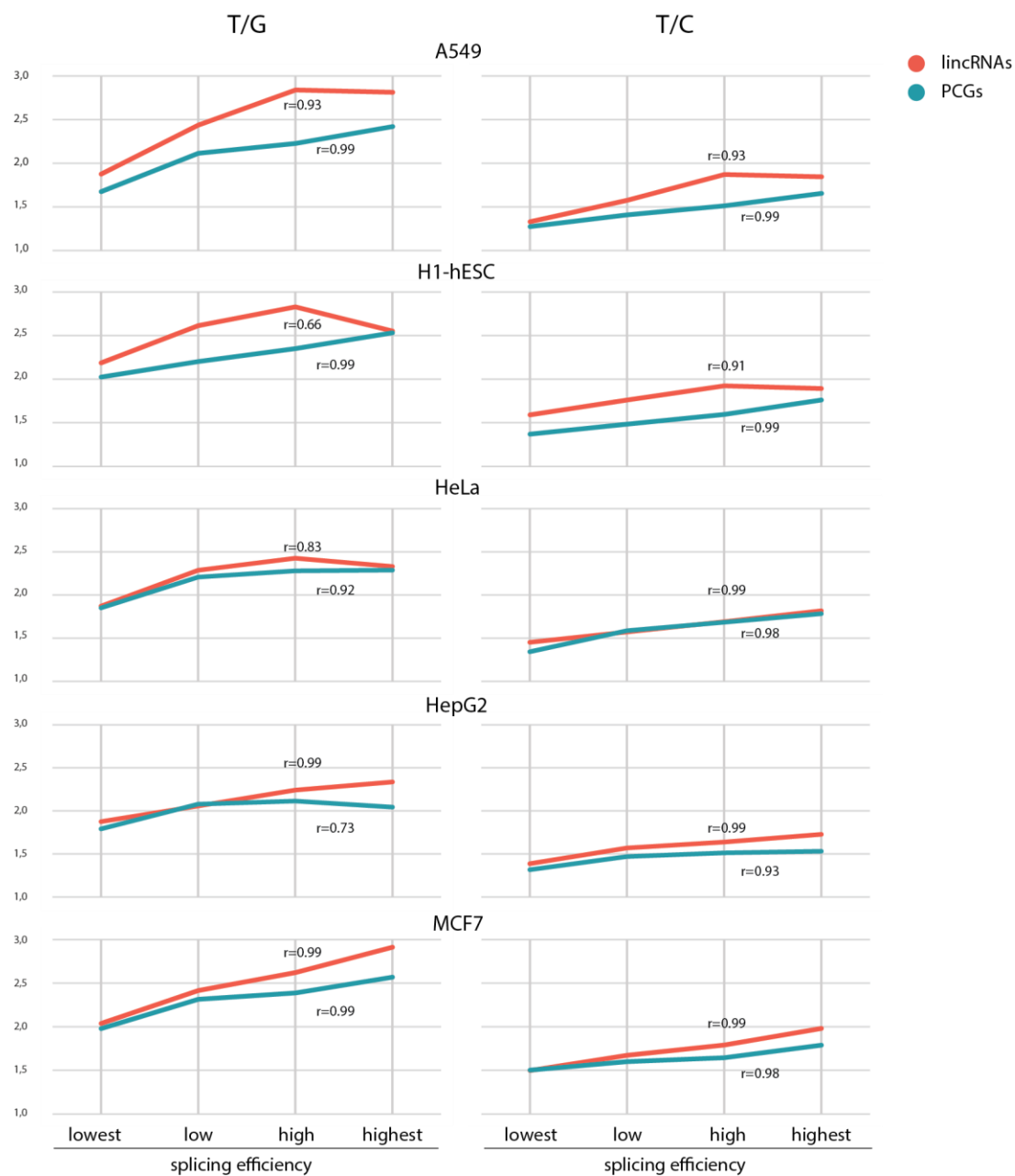


Figure 19. Thymidines over guanidines (T/G) or cytidines (T/C) ratio frequencies in the PPT region in different cell lines divided according to splicing efficiency in lincRNAs and PCGs; r - Pearson's correlation coefficient between T content in PPT and splicing index. Splicing groups are the same as at Figure 17.

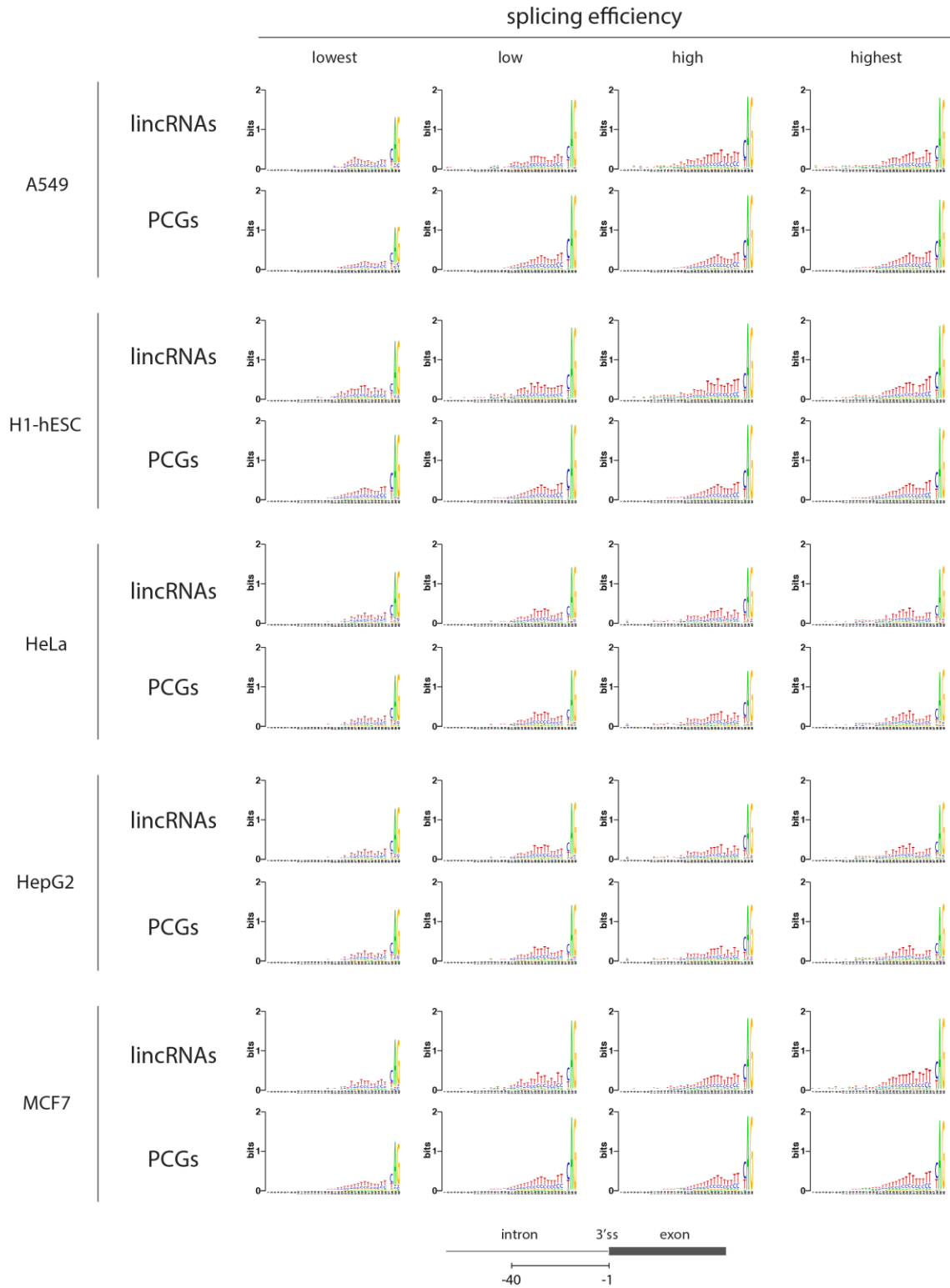


Figure 20. Sequence logos in 3'ss and PPT regions in lincRNAs and PCGs in different cell lines divided according to splicing efficiency.

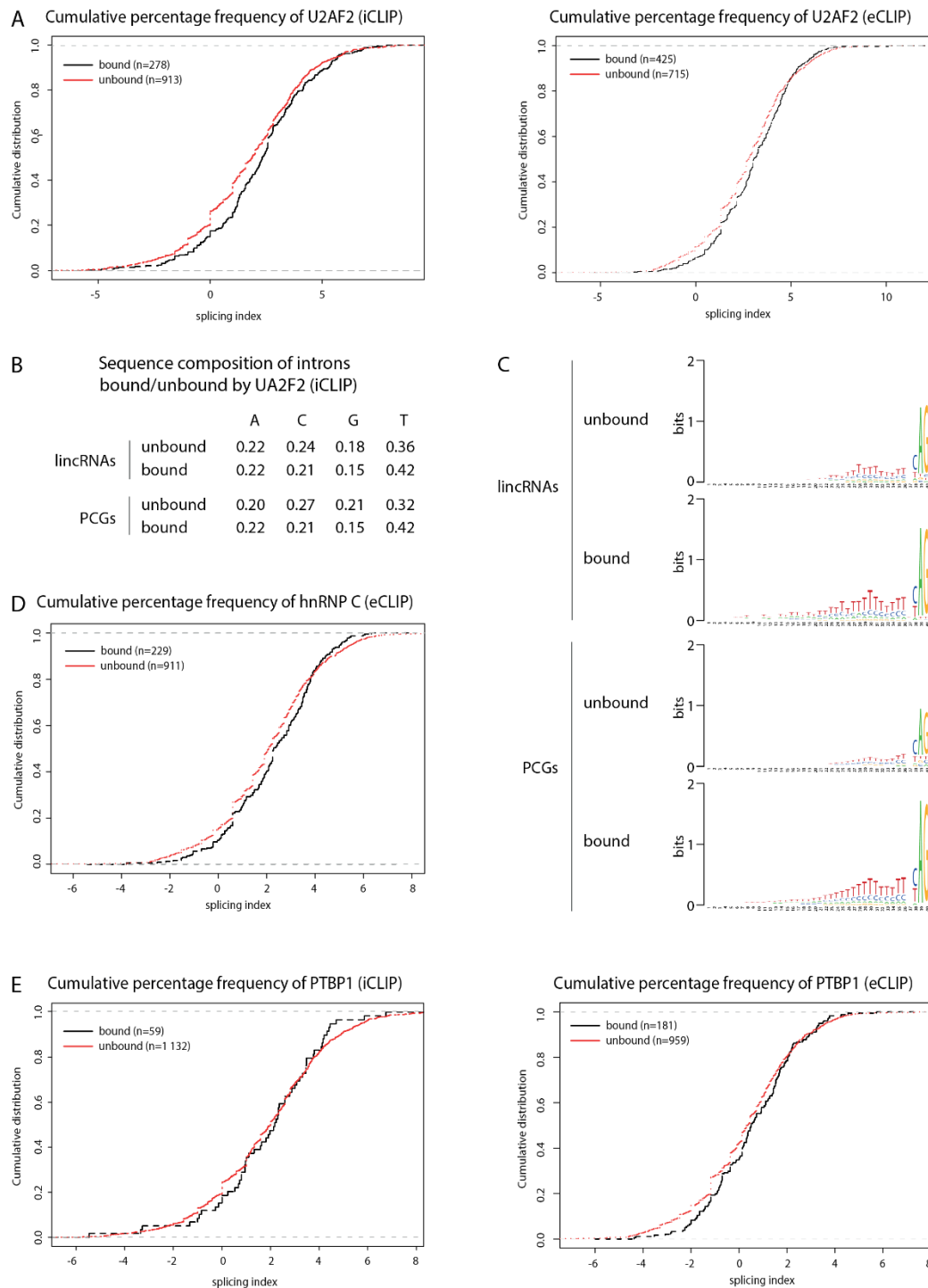


Figure 21. Cumulative distribution of splicing efficiency (A), sequence composition (B) and logos of 40 nt upstream of 3'ss (C) of lincRNA introns bound and unbound by U2AF2. Cumulative distribution of splicing efficiency comparing hnRNP C-bound and unbound lincRNAs (D), and PTBP1-bound and unbound lincRNAs (E); n=number of genes. PCGs serves as a control.

SR proteins bind less efficiently to lincRNAs

Our data suggest that a strong 5'ss and T-rich PPT together with productive U2AF2 binding are required for efficient lincRNA splicing. To assess whether additional factors may affect lincRNA splicing, we analyzed cumulative lengths of exons and introns and found that lincRNAs contain slightly longer introns and exons than PCGs (**Figure 22A**).

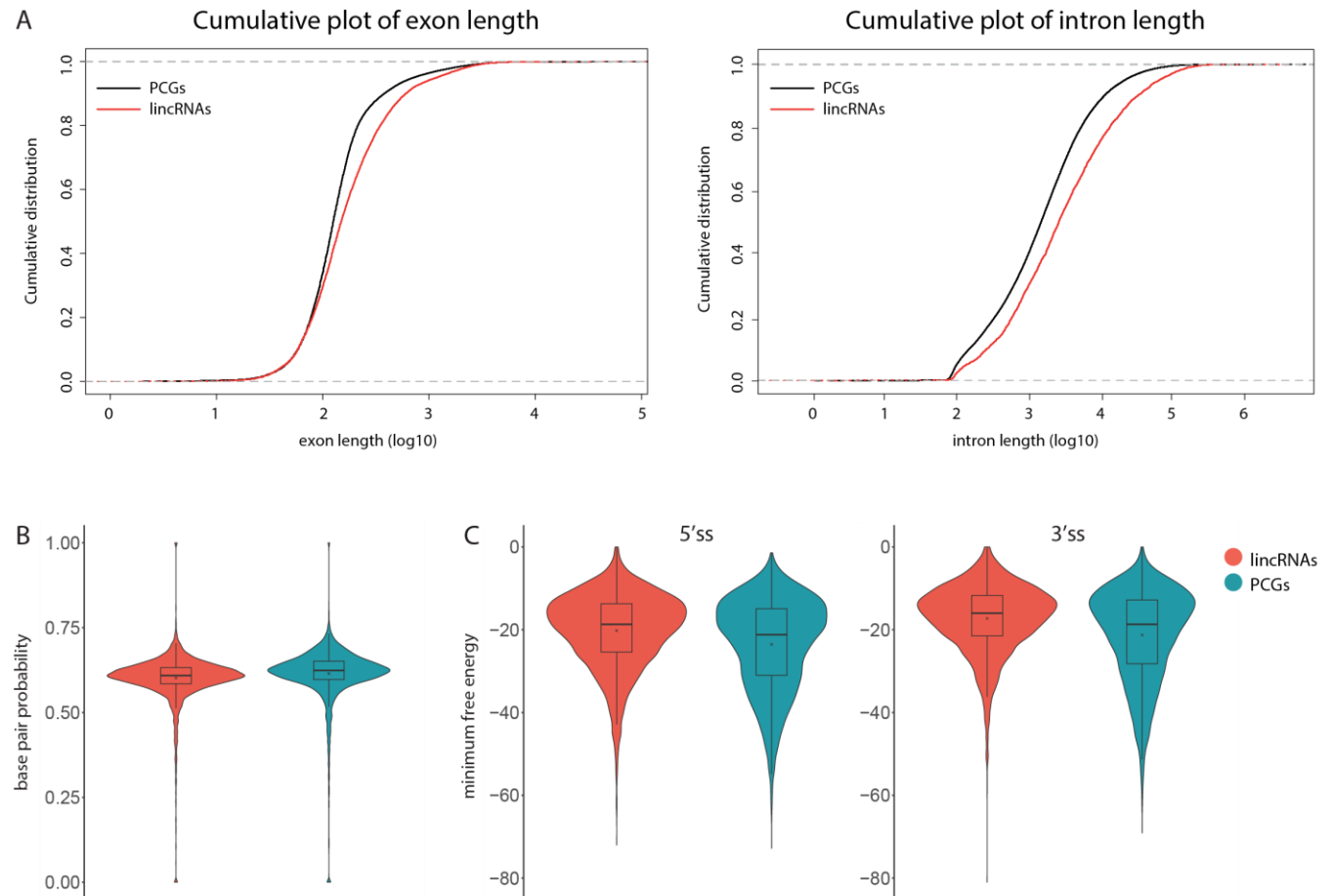


Figure 22. Comparative analysis of various properties between lincRNAs and Protein-Coding Genes (PCGs) (A) Cumulative distributions of exon and intron lengths in lincRNAs and PCGs. (B) Base pair probability of 5000 randomly chosen lincRNAs and PCGs shorter than 2000 nts. (C) Minimum free energies of RNA sequences 100 nt upstream and 100 nt downstream of 5'ss (left) and 3'ss (right).

5000 randomly chosen introns from lincRNAs and PCGs were analyzed. Because the length of introns has been associated with splicing efficiency (Klinz and Gallwitz 1985; Sterner et al. 1996; Bell et al. 1998; Fox-Walsh et al. 2005; Dewey et al. 2006; Louloui et al. 2018), longer introns can partially explain lower splicing efficiency of lincRNAs. We also calculated a putative base-pairing and minimum free energy of 5,000 randomly selected introns and did not find any relevant differences between lincRNAs and PCGs, which suggest that RNA secondary structure is not the major factor that would determine the splicing difference between PCGs and lincRNAs (**Figure 22B-C**). Finally, we analyzed the presence of

known splicing inhibitory sequences 100bp upstream and downstream of 3'ss. We found only one (out of 12) inhibitory motif (GTAGGT) enriched in lincRNAs over PCGs (**Figures 23 and 24**). Together these results indicate that except longer introns, lincRNAs do not contain any particular feature that would specifically inhibit their splicing.

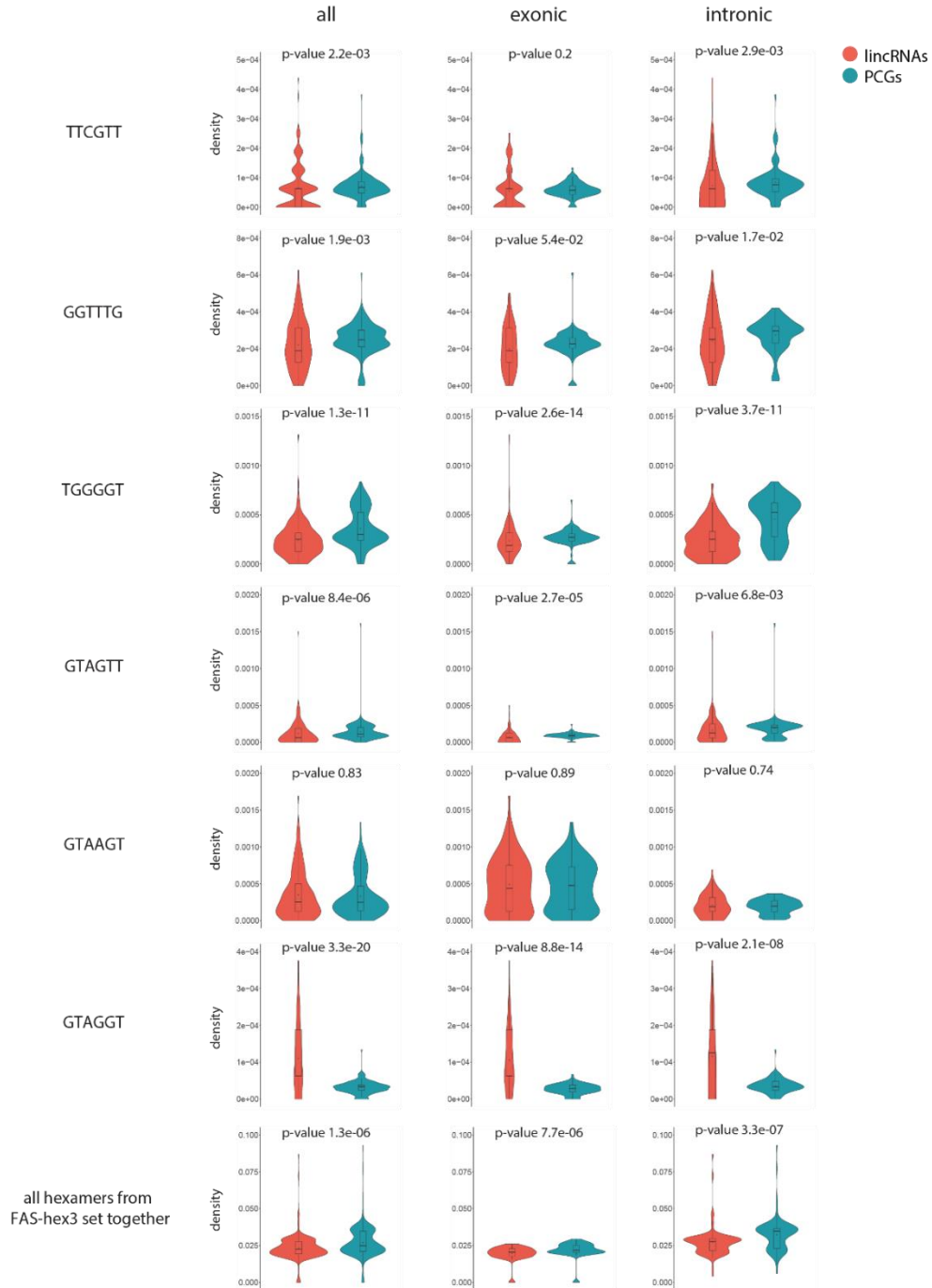


Figure 23. Splicing silencer motif distributions 100 nt upstream (intronic) and 100 nt downstream (exonic) of 3'ss in lincRNAs and PCGs. Motifs taken from Wang et al. (2004) and FAS-ESS web server (<http://genes.mit.edu/fas-ess/>). P-values are calculated by Wilcoxon rank sum test.

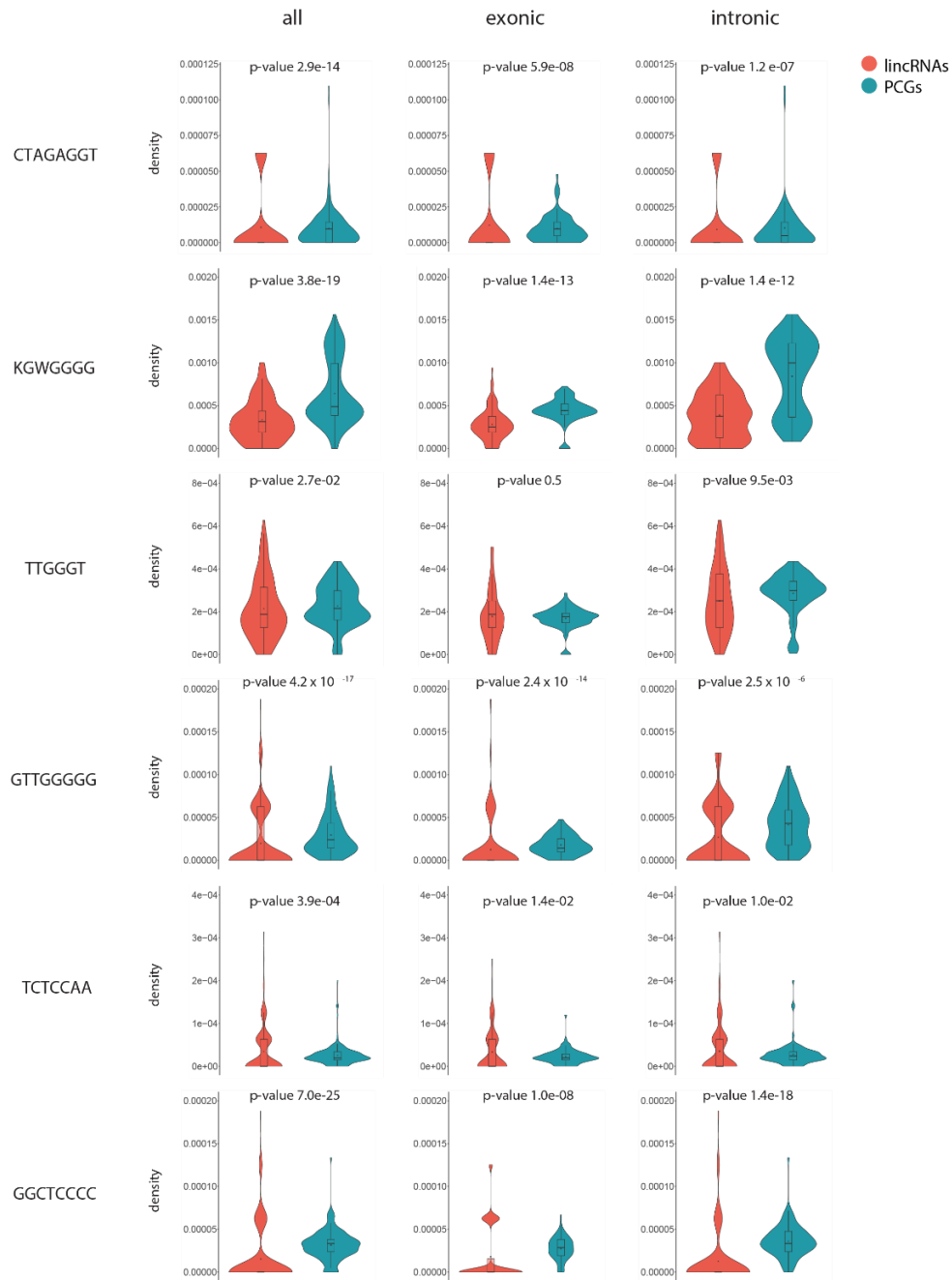


Figure 24. Splicing silencer motif distributions 100 nt upstream (intronic) and 100 nt downstream (exonic) of 3'ss in lincRNAs and PCGs. Motifs taken from Sironi et al. (2004). P-values are calculated by Wilcoxon rank sum test.

High dependence on strong splice sites could signal that lincRNAs lack additional splicing enhancer sequences that navigate the basic splicing machinery to splice sites. However, previous bioinformatic analyses showed that the global density of exonic splicing enhancers (ESEs) is even slightly higher in lincRNAs than in PCGs (Mele et al. 2017). In addition, ESEs are conserved in

lincRNAs, and no difference in the number of ESEs has been observed between efficiently and inefficiently spliced lincRNAs (Schüler et al. 2014; Haerty and Ponting 2015; Mele et al. 2017). To perform a more focused analysis, we searched for the occurrence of ESE motifs that are known to be recognized by SR proteins, general splicing enhancers (Paz et al. 2010; Mueller and Hertel 2011; Müller-McNicoll et al. 2016). We determined the occurrence of 28 ESE consensus motifs in exons and observed a striking difference in motif densities between lincRNAs and PCGs (Figure 25A and Figures 26-27). Only one motif (SRSF3 - WCWWC) was significantly enriched in lincRNA exons while the majority of analyzed SR binding motifs were more prevalent in PCGs (Figure 25A).

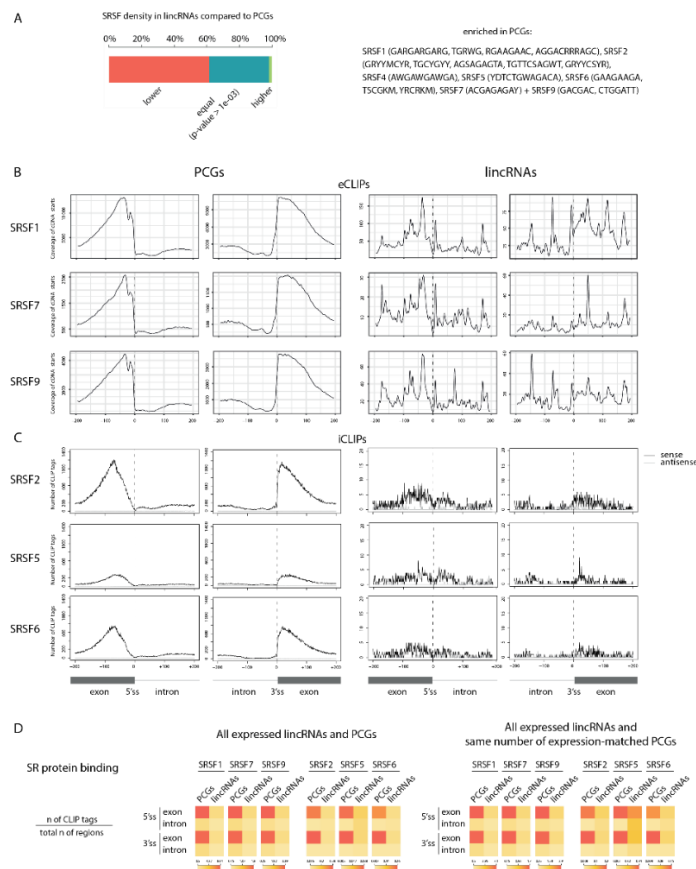


Figure 25. SR proteins preferentially bind to exons of PCGs. (A) Combined distribution of 28 SR protein binding motifs within 100 nt regions upstream of 5'ss and downstream of 3'ss between lincRNAs and PCGs. For detail distribution see Figures FIGURE 25 and FIGURE 26. Motifs enriched in PCG exons are indicated in left. (B) Binding of SRSF1, SRSF7, and SRSF9 within 200 nt regions around 5'ss and 3'ss in lincRNAs and PCGs was analyzed using available eCLIP data (BioRxiv: <https://doi.org/10.1101/179648>). Protein binding per region 200bp upstream/downstream of splice sites are shown for PCGs and lincRNAs.

(C) Binding of SRSF2, SRSF5, and SRSF6 within 200 nt regions around 5'ss and 3'ss in lincRNAs and PCGs was determined by iCLIP. Protein binding per region 200bp upstream/downstream of splice sites are shown for PCGs and lincRNAs. (D) eCLIP and iCLIP tags (FDR<0.05) in 200 nt regions upstream and downstream of 5'ss or 3'ss were selected, and the iCLIP tags per nucleotide divided by the total number of expressed regions are shown. Data are shown for all expressed lincRNAs, and PCGs (left diagrams) or all expressed lincRNAs and the identical number expression-matched PCGs (right diagrams).

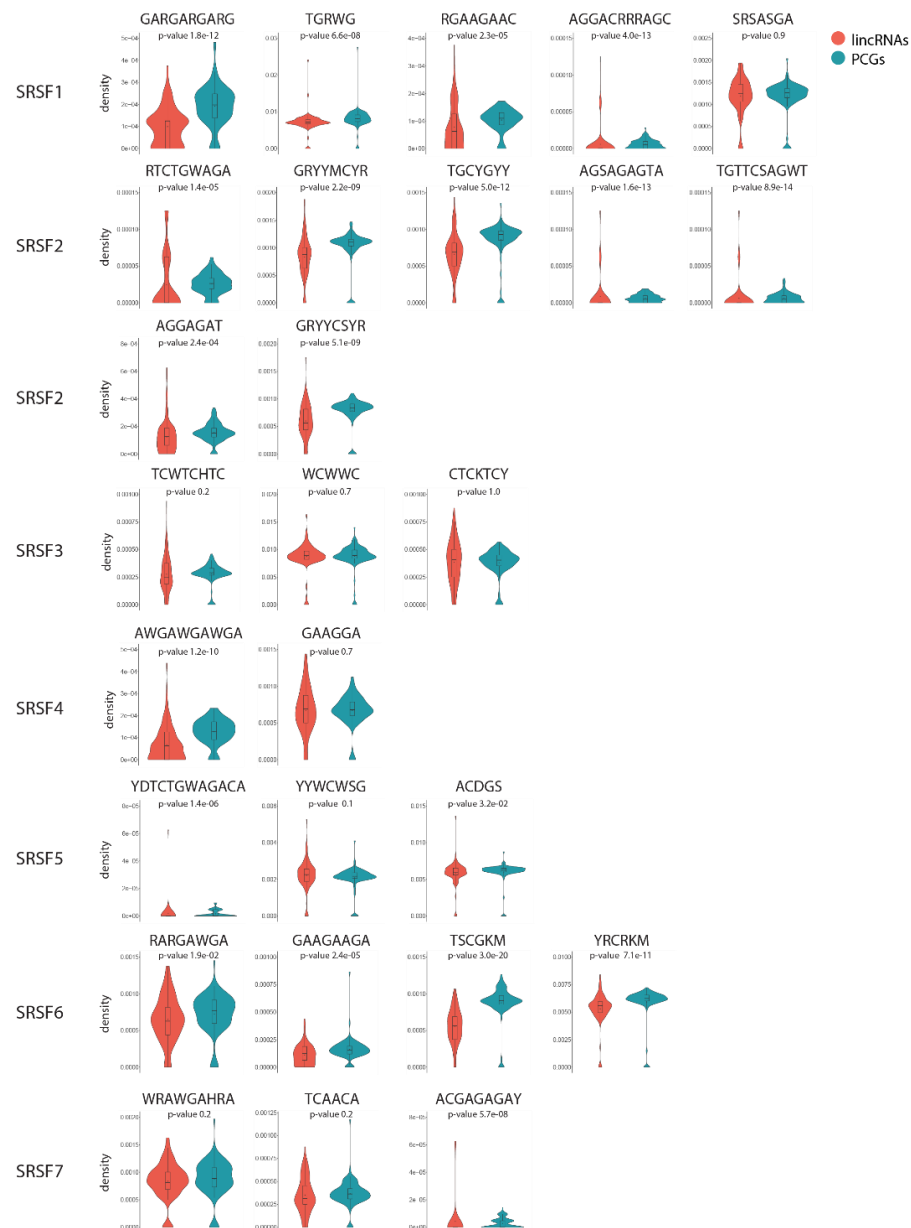


Figure 26. SRSF binding motif distributions in lincRNAs and PCGs 100 nt upstream of 5'ss. References for motifs are provided in Supplementary Data. P-values are calculated by Wilcoxon rank sum test. Two binding motifs of SRSF9 are not shown because no motifs were found in lincRNAs.

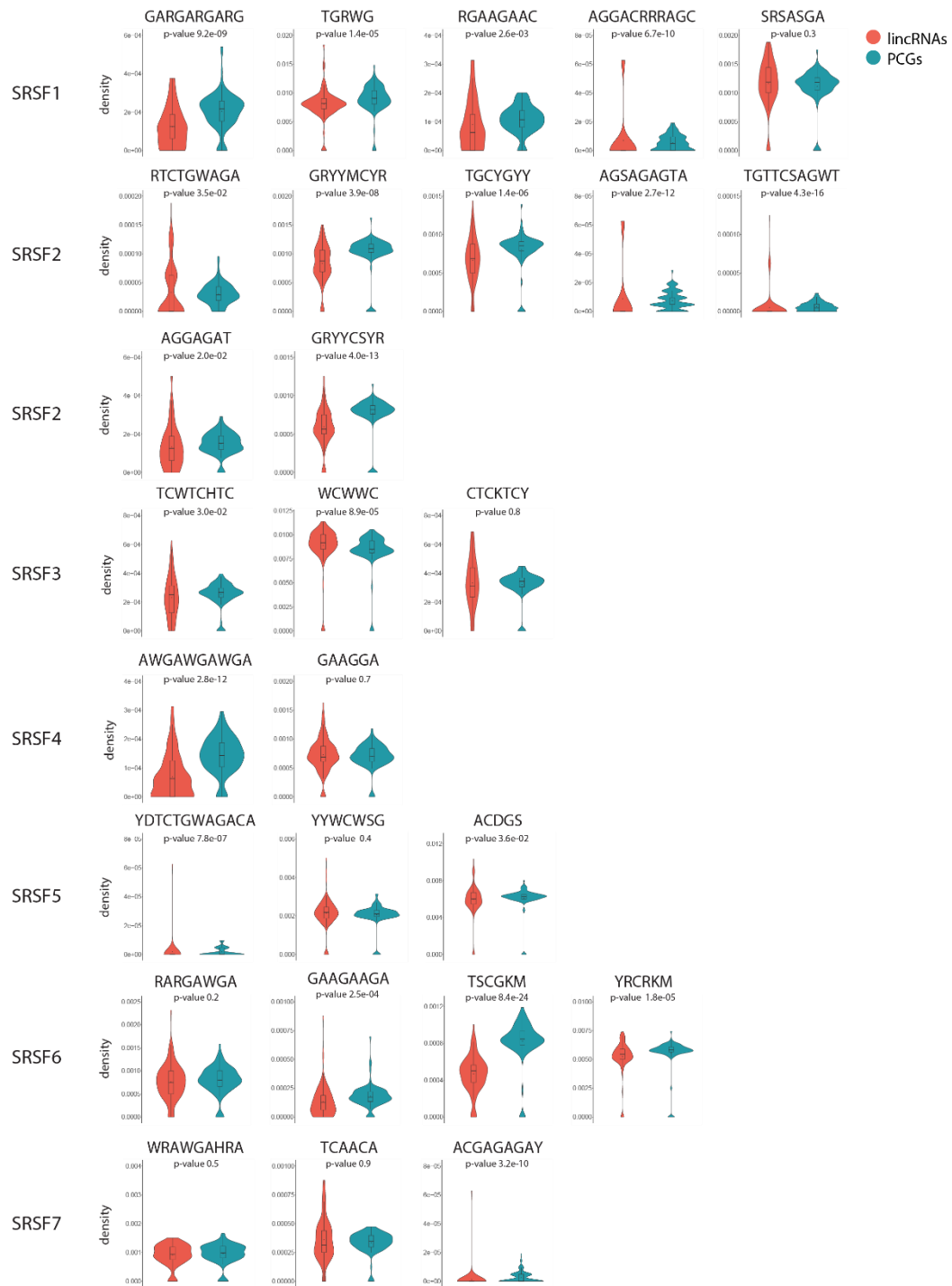


Figure 27. SRSF binding motif distributions in lincRNAs and PCGs 100 nt downstream of 3'ss. References for motifs are provided in Supplementary Data. P-values are calculated by Wilcoxon rank sum test. Two binding motifs of SRSF9 are not shown because no motifs were found in lincRNAs.

To test whether a smaller number of SR binding motifs in lincRNAs results in a lower interaction with SR proteins, we analyzed available eCLIP data performed with SRSF1, 7 and 9 in HepG2 cells (BioRxiv: <https://doi.org/10.1101/179648>). All three SR proteins bound efficiently within PCG exons, while their association with lincRNAs was much weaker and we did not detect any significant enrichment over exons (**Figure 25B**). SR protein binding to lincRNAs was lower compared to the total expressed PCGs (18-26% of binding to PCGs; **Figure 25D**, left panel). To normalize for the expression level of PCGs and lincRNAs, we created a subset of PCGs that match number and expression level of lincRNAs and repeated the analysis. Similarly, binding of SR proteins to lincRNAs was reduced to 20-30% of expression-matched PCGs (**Figure 25D**, right panel). To investigate the binding of additional SR proteins not covered by eCLIP, we performed iCLIP in HeLa cell lines stably expressing GFP-tagged SRSF2, SRSF5 or SRSF6 from bacterial artificial chromosomes at near endogenous levels using anti-GFP antibodies as described before (Botti et al. 2017). iCLIP libraries were prepared in triplicates, submitted to deep sequencing, and significant cross-link events of individual SR proteins were identified (Botti et al. 2017). Similarly to previous studies (Fairbrother et al. 2004; Xiao et al. 2007), SR proteins bound preferentially to exonic regions of PCGs (**Figure 25C**). In agreement with the eCLIP data, binding of all three analyzed SR proteins to lincRNAs was much lower compared to all expressed PCGs (13-30% of binding to PCGs; **Figure 25D**, left panel) or expression-matched PCGs (56-68% of binding to PCGs; **Figure 25D**, right panel). Altogether, this confirmed that SR proteins interact poorly with lincRNAs, which is independent of lincRNA expression level.

To test whether residual SR protein binding nevertheless promotes splicing of lincRNAs, we compared splicing efficiencies of bound and unbound lincRNAs (using PCGs as control). Analysis of lincRNA interactions with the three SR proteins identified by iCLIP revealed that binding of SR proteins in lincRNA exons (100 nt downstream of 3'ss) improves their splicing efficiencies (**Figure 28A**). However, no such correlation was found for SR protein interactions with lincRNAs identified by eCLIP (**Figure 28C**), which might be due to the higher noise in eCLIP data compared to iCLIP data (Haberman et al. 2017). Interestingly, from iCLIP data, the stimulatory effect of SR protein binding on splicing of PCGs appears much weaker than that of lincRNAs (**Figure 28B**), in contrast to eCLIP data, where the stimulatory effect appears quite strong (**Figure 28D**). The binding of analyzed SR proteins in intronic regions (100 nt upstream of 3'ss) (**Figure 28B**) correlates with lower splicing efficiency, which is consistent with the proposed position-dependent

splicing activity of SR proteins (Kanopka et al. 1996; Gallego et al. 1997; Jiang et al. 1998; ten Dam et al. 2000; Simard and Chabot 2002; Wang et al. 2005; Buratti et al. 2007; Erkelenz et al. 2013).

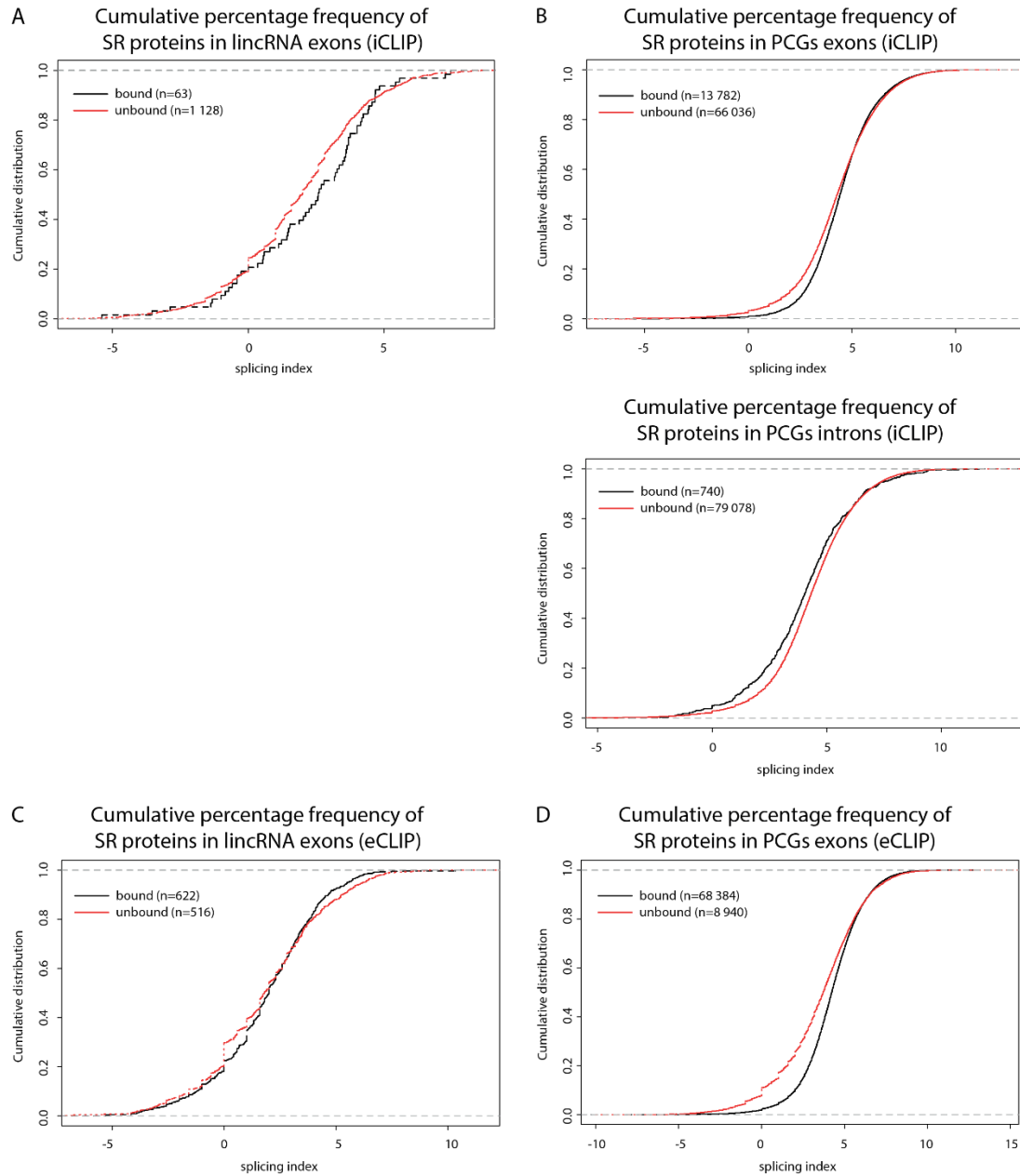


Figure 28. Cumulative distributions of splicing efficiencies comparing SRSF2, SRSF5 and SRSF6-bound and unbound lincRNAs (A) and PCGs (B) in exons (100 nt downstream of 3'ss) and introns (100 nt upstream of 3'ss). Cumulative distributions of splicing efficiencies comparing SRSF1, SRSF7 and SRSF9-bound and unbound lincRNAs (C) and PCGs (D) in exons (100 nt downstream of 3'ss); n = number of genes.

7. Results - Project 2

Canonical and non-canonical introns in diplonemid genomes

Draft genome assemblies were obtained for *A. motanka*, *N. karyoxenos*, *S. specki*, *L. lanifica*, *D. japonicum*, and *R. humris*. The genome assembly sizes ranged between 35 Mb and 118 Mb for *S. specki* and *N. karyoxenos*, respectively, and N₅₀ values varied between 1,160 bp for *N. karyoxenos* and 11,390 bp for *L. lanifica* (**Table 1**). Next, we performed intron prediction aided by mapping RNA-seq data to the genome assemblies (see **Figure 29** and Materials and Methods for details). To ensure the acquisition of high-confidence intron models, we excluded introns supported by fewer than 15 split reads. This filtering process yielded 98,221 introns from *A. motanka*, 21,076 from *N. karyoxenos*, 60,346 from *S. specki*, 12,835 from *L. lanifica*, 10,131 from *D. japonicum*, and 13,462 from *R. humris*.

Table 1. Genome assembly statistics.

	No. Contig	Total length	Max length	N50 [#]	N90 [#]	N80 [#]
<i>A. motanka</i>	91695	81481355	34661	2658	231	957
<i>N. karyoxenos</i>	102844	118343072	124812	1160	770	834
<i>S. specki</i>	35642	34908406	19484	1175	757	833
<i>L. lanifica</i>	15958	73316053	83793	11390	2873	4798
<i>D. japonica</i>	160587	84236011	293136	2515	144	205
<i>R. humris</i>	189683	87108244	57896	2226	141	184

[#]The N50 value is the length of the shortest contig that covers 50% of the genome, while N80 and N90 are the contig lengths at 80% and 90% of the genome, respectively.

Next, we aimed to identify and characterize common intronic motifs across these six diplonemid species. We categorized the introns based on their splice site sequences and those possessing GT and AG dinucleotides at their 5' and 3' splice site, respectively, classified as canonical introns. Conversely, introns with the splice sites deviating from the canonical 'GT-AG' motif were classified as non-canonical introns. We observed an overwhelming dominance of canonical introns in *S. specki* (98.44%), *L. lanifica* (97.45%), *D. japonicum* (98.33%) and *R. humris* (97.10%), reaffirming the prevailing trend observed in the wide majority of eukaryotes.

However, the picture was very different in two Hemistasiidae species, *A. motanka* and *N. karyoxenos*, as these diplonemids turned out to be extremely rich in the non-canonical splice sites, which constituted 96.93% and 96.644% of all splice sites, respectively (**Figure 30**). Further

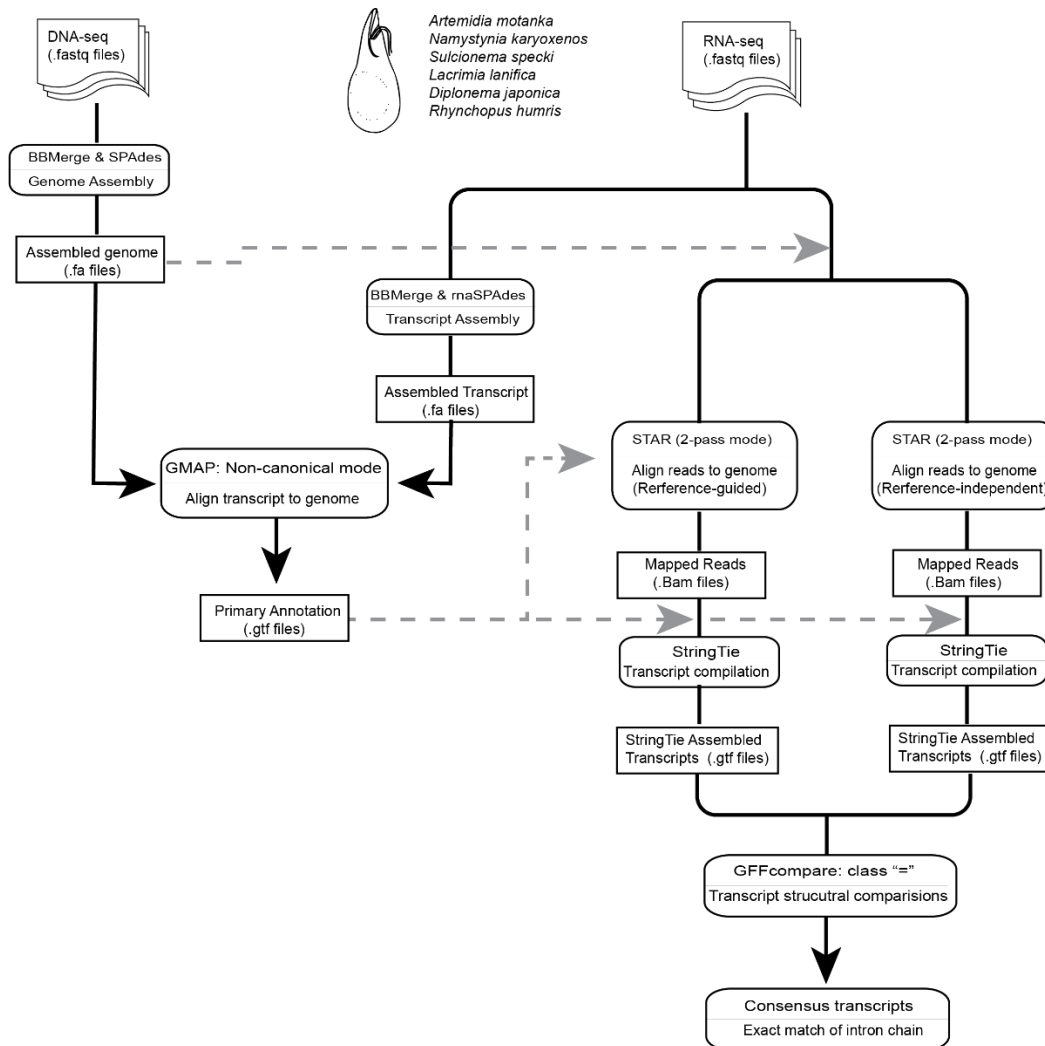


Figure 29. Intron prediction pipeline.

analysis of these non-canonical splice site categories revealed a diverse landscape of dinucleotides at the intron ends. Most common combinations were CT-CG, TG-GC, CG-AG and GC-CA, each amounting to less than 12% of all the identified intron ends (**Figures 30 and 31**), which indicates a lack of splice site conservation.

In the following step, we have mapped sequences involved in intron definition and removal, namely the 5' splice site (5'ss) and the polypyrimidine tract/3' splice site (PPT-3'ss) (**Figure 30**). The primary aim of this analysis was to uncover the consensus sequence and the

extent of conservation in the proximity of the splice sites. The 5'ss included 3 bases at the 3' end of upstream exon and 6 bases at the 5' end of intron, while the 3'ss was defined as 20 bases in the intron and 3 bases in the downstream exon. In *L. lanifica*, *D. japonicum* and *R. humris*, the most frequent nucleotides were (G)/GTATG at the 5'ss, while only a shorter G/GT motif is present in *S. specki* (**Figure 30**). The 5'ss region was strikingly different in *A. motanka* and *N. karyoxenos* and displayed a strong preference for C and G nucleotides (**Figure 30**). Furthermore, we observed a preference for C in the pyrimidine track of each species, except for *D. japonicum*, where, unusually, A was the most abundant nucleotide. Weak conservation of A in the PPT-3'ss was also found in *L. lanifica* and *R. humris*. *S. specki* exhibited the shortest PPT primarily comprised of pyrimidine (C and T) residues.

Finally, we analyzed the 3' end of introns. Four Diplonemidae species have a conserved CAG sequence in this site, while there is no strong conservation at the 3' ends of introns of the two examined representatives of Hemistasiidae. Combined, these data suggest that intronic splicing signals in diplonemids are diverse and species-specific, with particularly distinct non-canonical intron-defining patterns in *A. motanka* and *N. karyoxenos*.

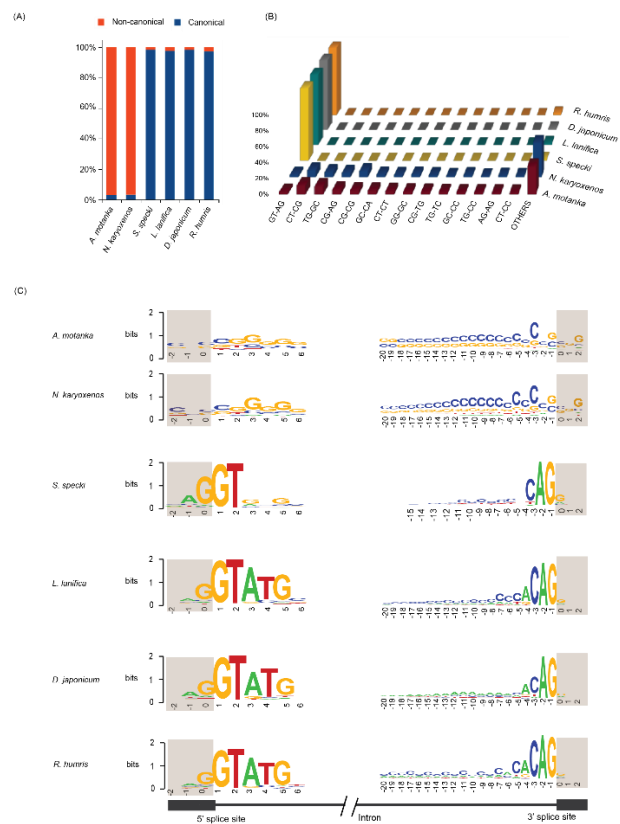


Figure 30. Splice-site sequences in diplonemids. (A) Stacked bar plot shows the percentage of canonical and non-canonical introns. (B) Distribution of splice site sequences. First two nucleotides from 5' end and last two nucleotides from 3' end of introns are shown. (C) Sequence logos of 5' and 3' splice sites.

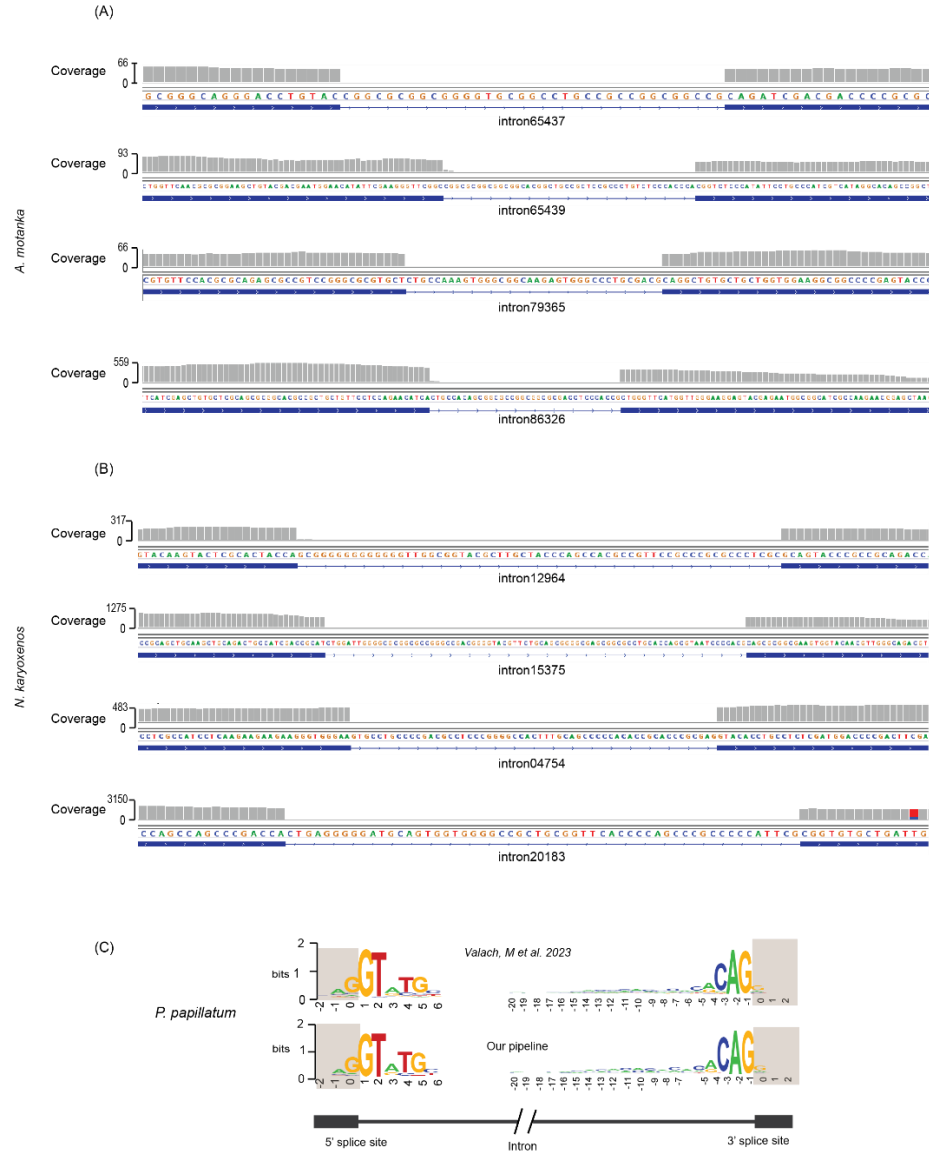


Figure 31. Sequence logos of splice sites: Visualization of RNA-seq read coverage and splice site annotations using IGV: (A) *A. motanka*: visualization of intron65437 (*contig08337:1450-1484*), intron65439 (*contig08337:1703-1748*), intron79365 (*contig048:5204-5239*), and intron86326 (*contig08012:2472-2507*). (B) *N. karyoxenos*: visualization of intron12964 (*contig050515:701-763*), intron15375 (*contig020064:1119-1199*), intron04754 (*contig07549:464-520*), and intron20183 (*contig015374:1065-1120*). The Integrated Genome Viewer (IGV) displays RNA-seq read coverage (top track) mapped to the genome alongside corresponding splice site annotations (bottom track). Exons are represented as thick blocks, while introns are depicted as thin connecting lines or gaps between the blocks in the annotation track. The coverage track highlights the read depth across exons, while intronic regions exhibit low or absent read coverage, consistent with splicing events. (C) *P. papillatum* splice site sequence logos from recently sequenced genome (Valach, M et al. 2023) and annotated introns using our pipeline

Characterization of the spliceosome components in diplonemids

The prevalence of non-canonical introns in Hemistasiidae lead us to investigate whether the splicing machinery differs among diplonemids. We identified all snRNAs in the assayed diplonemid genomes except in *R. humris* in which we could not find U1 snRNA. Based on homology, all identified snRNAs belong to the U2 type, rather than the U12 type (**Figures 32A, Figures 33-38**). Interestingly, *A. motanka* and *N. karyoxenos*, which contain a high proportion of the non-canonical introns (**Figure 30**), also encode well-conserved U1 snRNA genes that are complementary to canonical but not to non-canonical splice sites. However, we cannot rule out the possibility that *A. motanka* and *N. karyoxenos* contain also non-canonical U1 snRNAs, which could not be identified due to their extreme divergence or the incompleteness of the genome assemblies. We found three and two copies of U1 snRNA in the *D. japonicum* and *S. specki* genomes, respectively (**Figure 39A**). A more detailed inspection revealed that U1 snRNAs in *S. specki* differ in the 5' end sequence that base pairs with the 5'ss. One U1 snRNA contains a conserved sequence at the 5' end, which is replaced by a stretch of Cs in the other U1 snRNA gene (**Figure 39**). However, an examination of the potential base-pairing between the 5'ss of both the canonical and alternative U1 snRNAs showed comparable complementarity with the canonical 5'ss sequence logo identified in *S. specki* (**Figures 30 and Figure 39B**), suggesting a substantial flexibility and evolutionary divergence of its splicing machinery.

Search for the protein components of snRNPs led to the identification of 25 candidate orthologs, which represent almost all subunits found in mammals (**Figure 32B**)(Will and Lührmann 2011). Notably, we found that only ~55 amino acids at the N-terminus of the SNRPC (U1-C) protein that interact with U1 snRNA are conserved. These findings indicate that all species contain a functional set of snRNPs that catalyze removal of canonical introns.

Figure 4

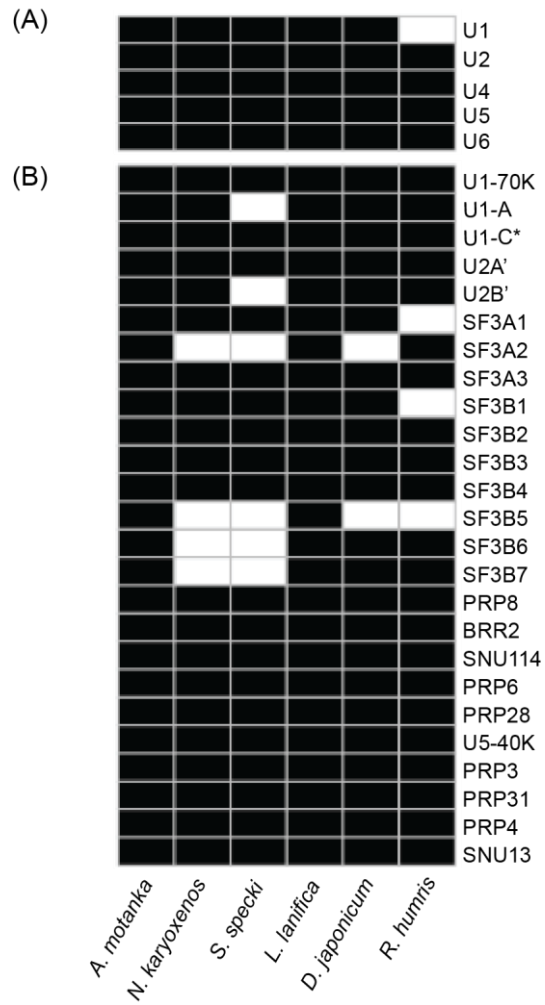


Figure 32. Identification of snRNP-associated spliceosomal components in diplonemids. (A) Identified snRNAs. **(B)** Identified orthologs of spliceosomal proteins. Black box indicates presence of the snRNA/protein. * indicates partial identified protein.

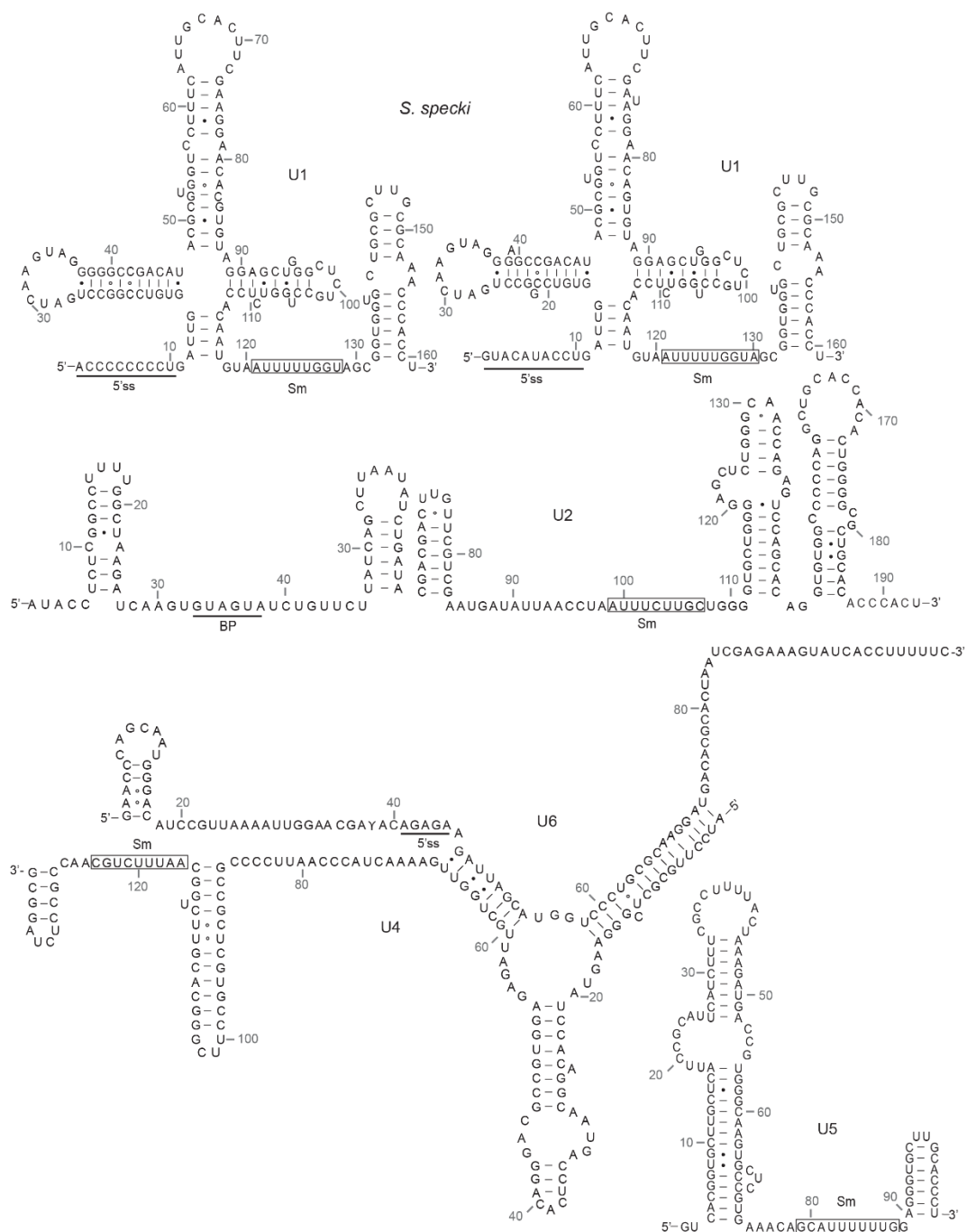


Figure 35. Secondary structure prediction for canonical U1, non-canonical U1, U2, U4, U5, and U6 from *S. specki*. 5'ss and BP binding regions are underlined, Sm-binding sites are indicated by black rectangles.

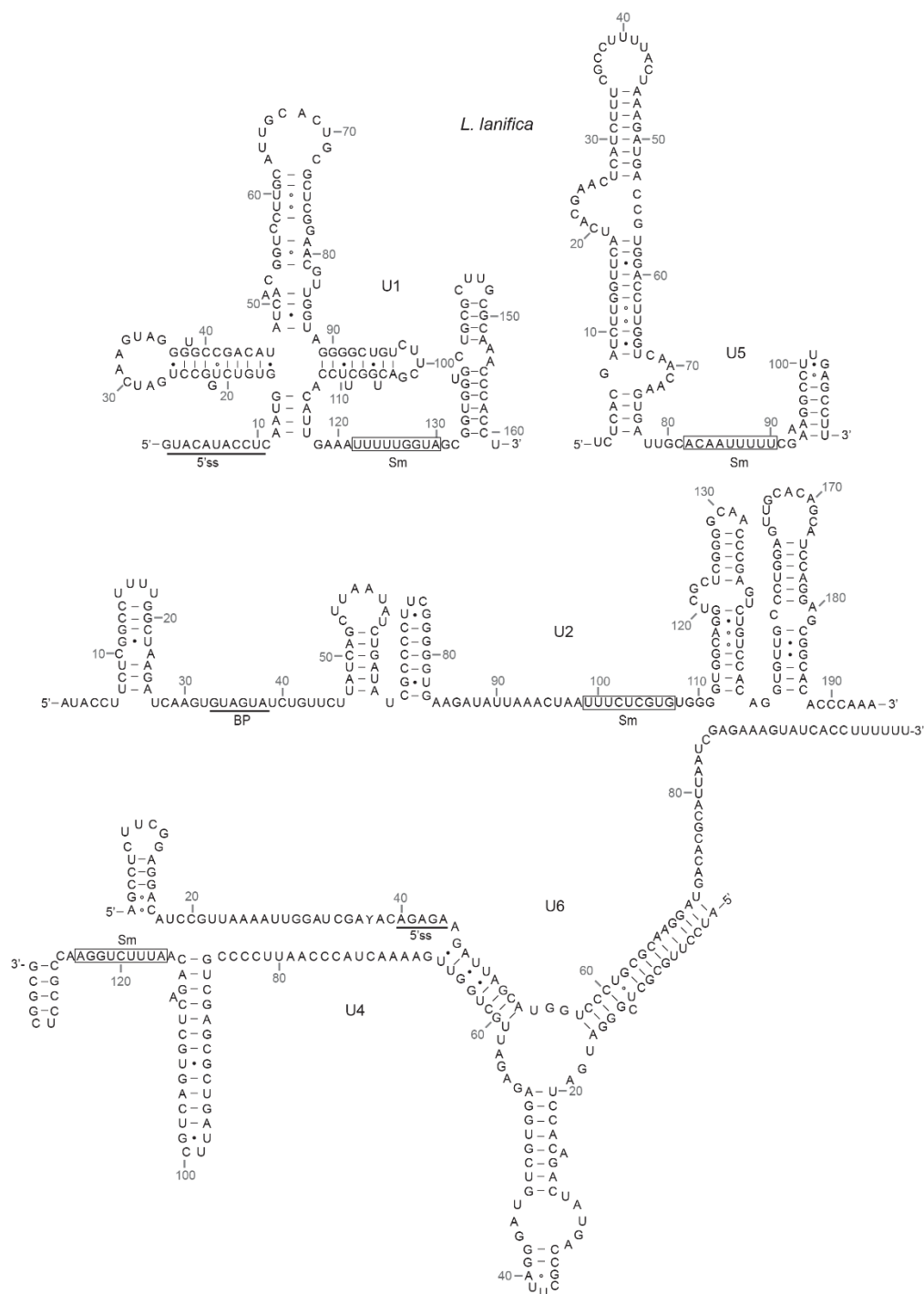


Figure 36. Secondary structure prediction for U1, U2, U4, U5, and U6 snRNAs from *L. lanifica*. 5'ss and BP binding regions are underlined, Sm-binding sites are indicated by black rectangles.

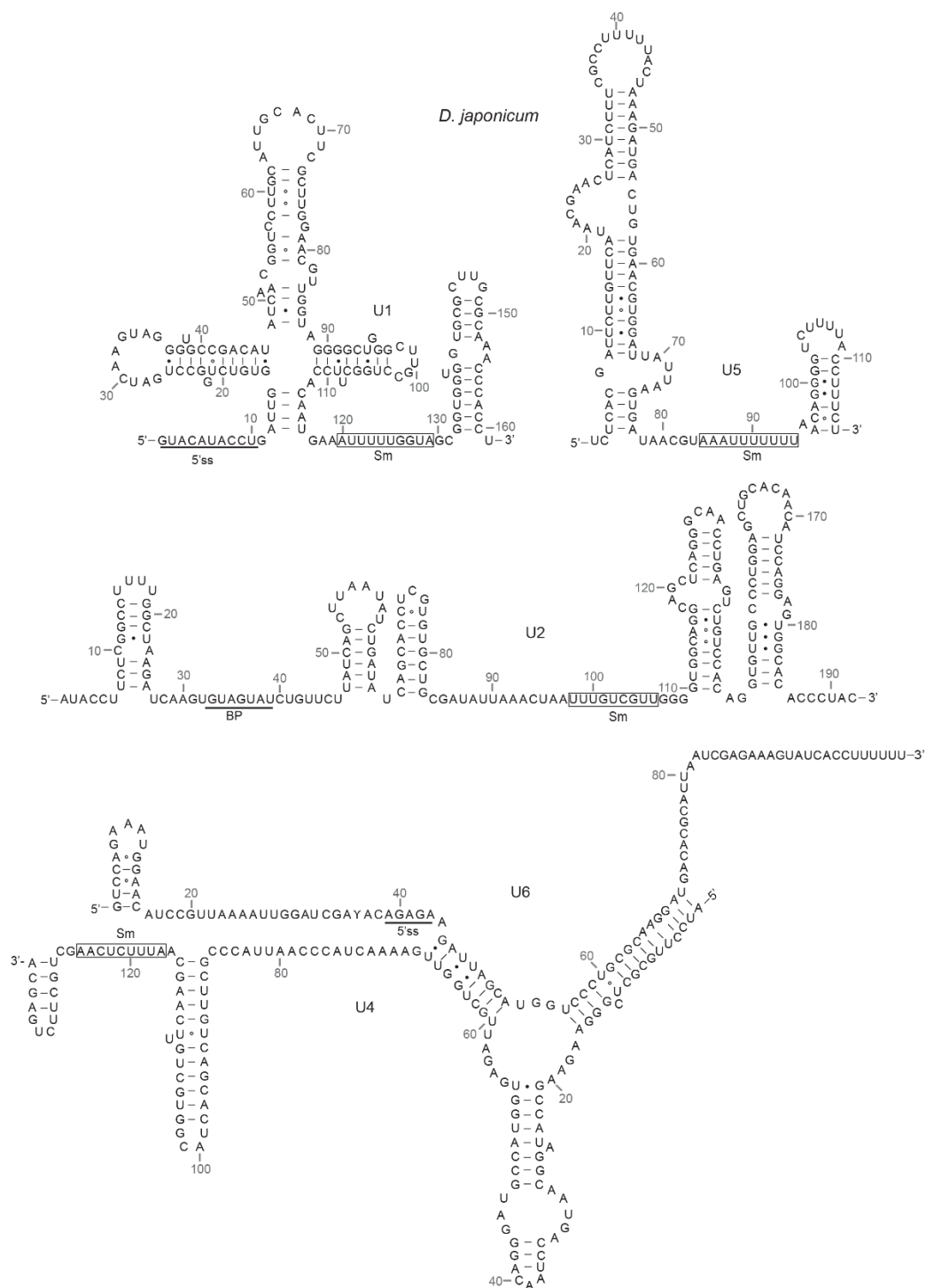


Figure 37. Secondary structure prediction for U1, U2, U4, U5, and U6 snRNAs from *D. japonicum*. 5'ss and BP binding regions are underlined, Sm-binding sites are indicated by black rectangles.

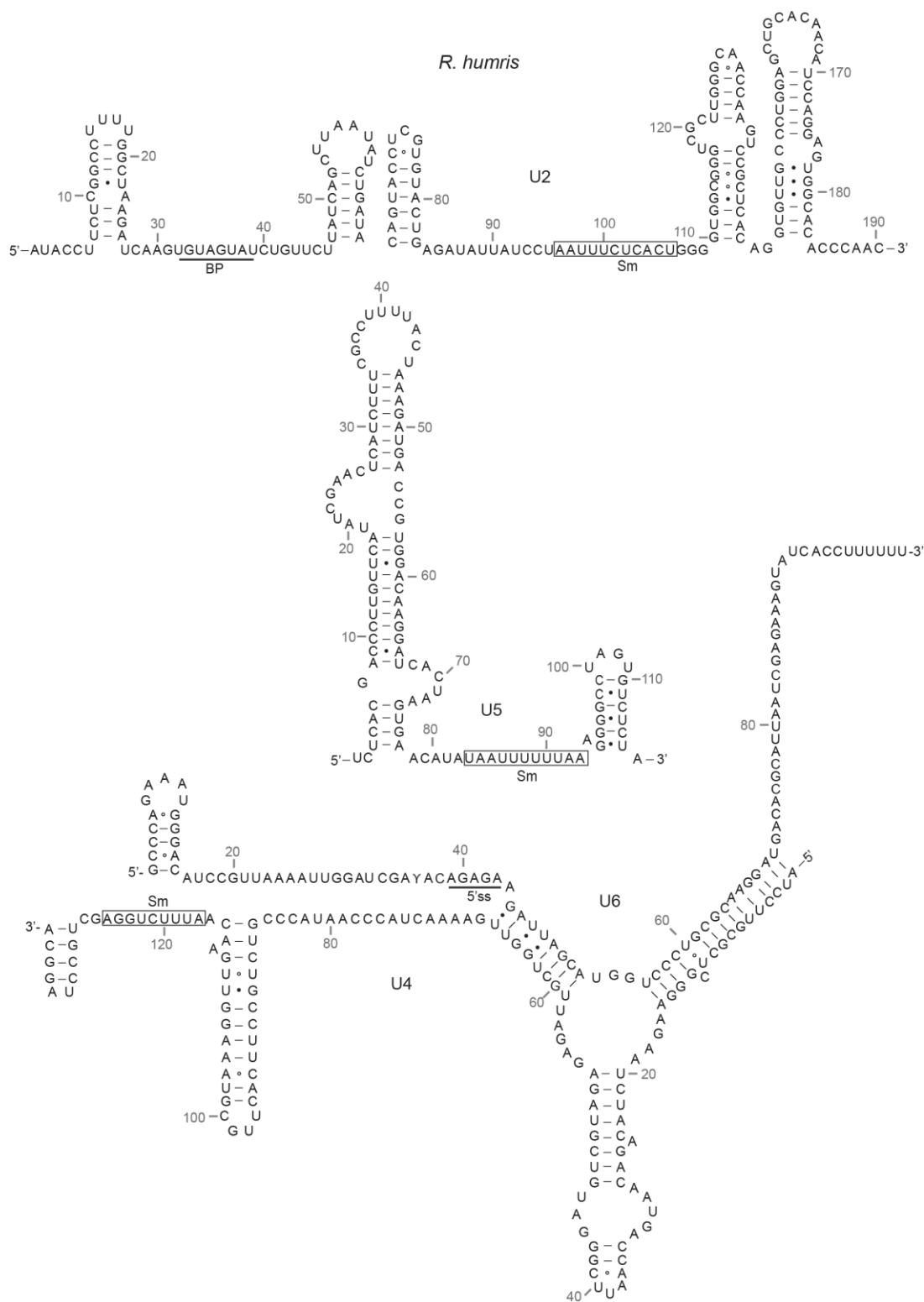


Figure 38. Secondary structure prediction for U2, U4, U5, and U6 for *R. humris*. The 5'ss and BP binding regions are underlined, Sm-binding sites are indicated by black rectangles.

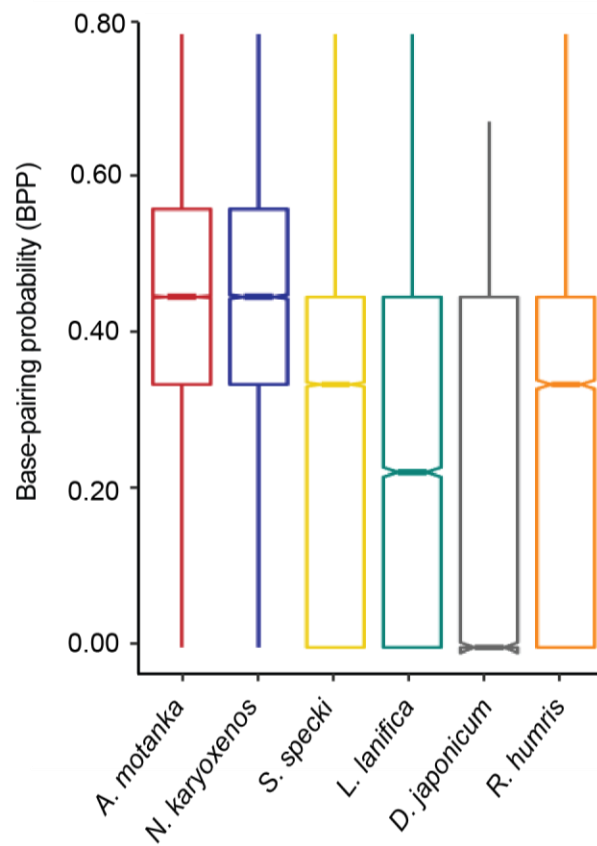


Figure 40. Base-pairing probability (BPP) between the 5'ss and 3'ss in various diptonemid species. The box plot displays the distribution of BPP values, with the box representing the interquartile range (IQR) between the first and third quartiles. The horizontal line within the box indicates the median BPP. The notches around the median represent the 95% confidence interval, providing a visual indication of the variability and significance of the median. BPP is calculated as the number of base-pairing nucleotides between 9 bases of the intronic 5'ss and 9 bases of the intronic 3'ss, divided by 18. ANOVA results indicate a significant difference in BPP among species ($\chi^2 = 6587$, $df = 5$, $p < 2e-16$).

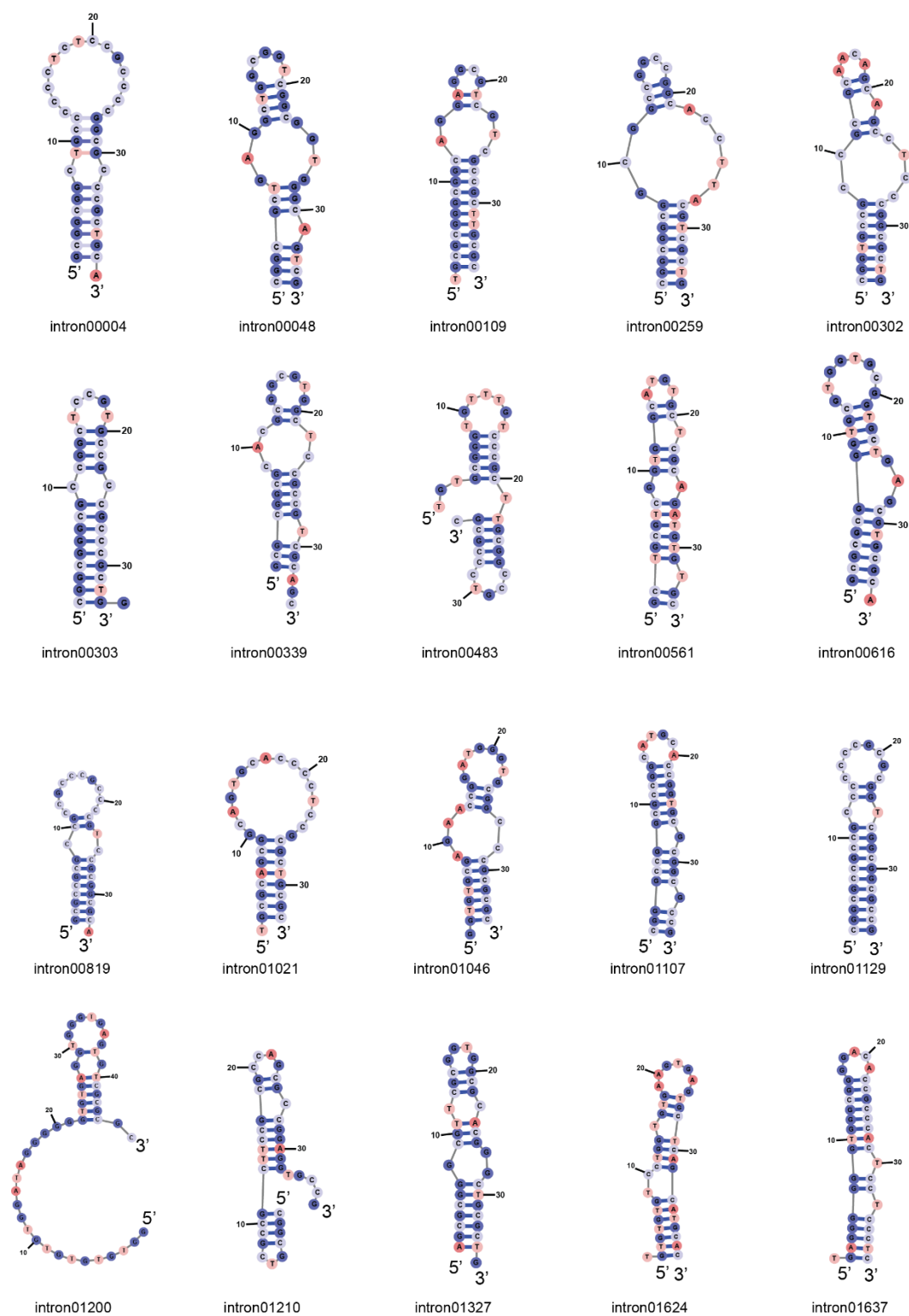


Figure 41. Secondary structure of selected introns from *A. motanka*.

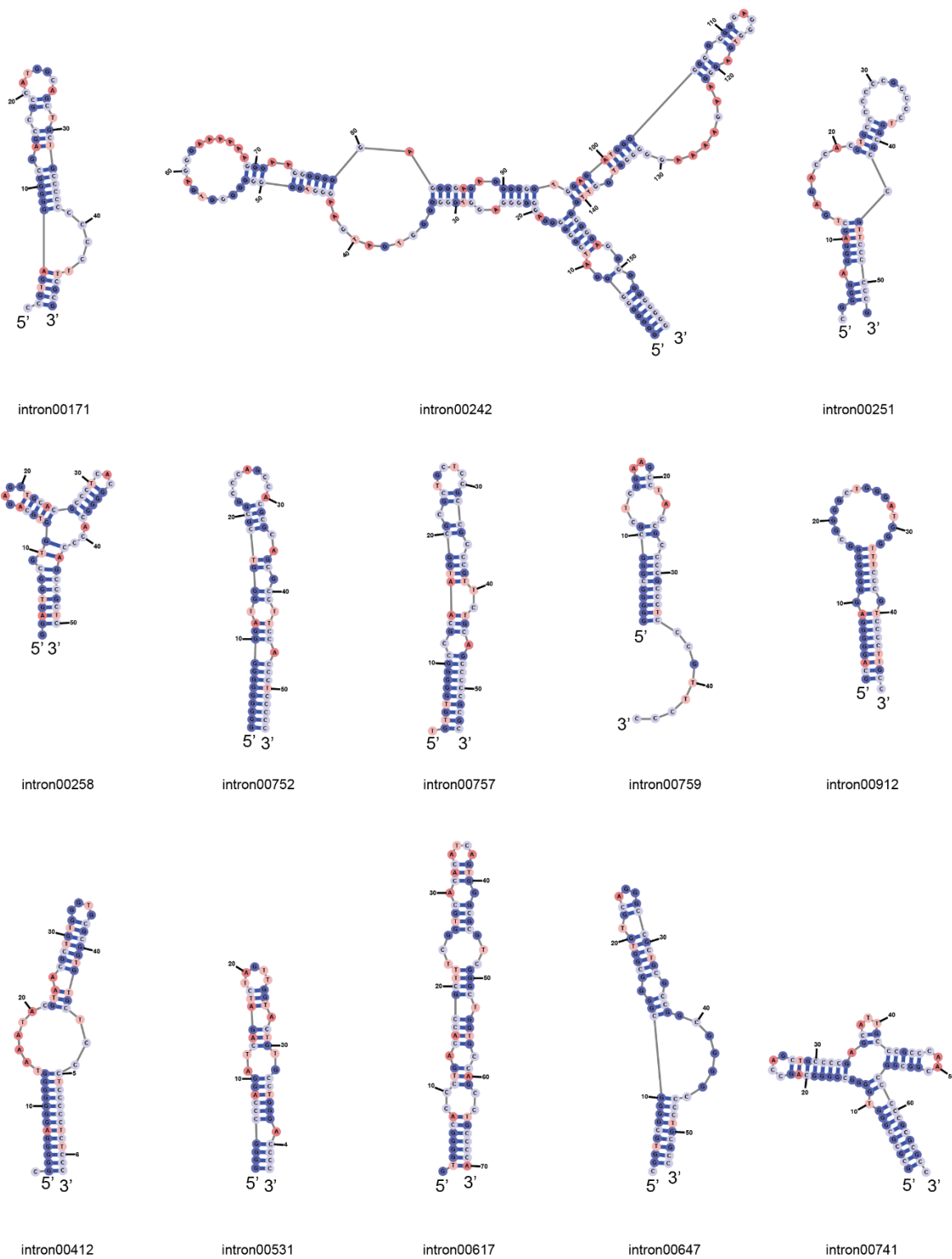


Figure 42. Secondary structure of selected introns from *N. karyoxenos*.

Splice site analysis by comparative genomics

To trace the potential origin of the non-canonical introns using a phylogenetic approach we compared the evolutionary trends of canonical and non-canonical splice sites of *A. motanka* and *N. karyoxenos* with other marine diplonemids including the recently sequenced *P. papillatum*, (Valach et al. 2023b). Moreover, to avoid masking signals from other splice site positions, conserved dinucleotides at the intron ends have been removed from the analysis. Position Weight Matrices (PWMs) for both the 5'ss and 3'ss were employed to compute a distance matrix between species. Resulting phylogenetic trees generated for both canonical and non-canonical splice sites revealed distinct clustering of the 5'ss and 3'ss, reflecting evolutionary relationships among these splice sites across species. Such a distinct clustering of canonical and non-canonical introns indicates that the non-canonical introns appeared in Hemistasiidae after the split of this family and Diplonemidae. Canonical and non-canonical splice sites of *A. motanka* and *N. karyoxenos* cluster together which suggest their common origin or less likely their convergent evolution (**Figures 43A and B**). Although *A. motanka* and *N. karyoxenos* are sister clades with long branches, mutual clustering of their canonical and non-canonical splice sites indicates co-evolution of these sequence elements. The branching order of diplonemids, so far based primarily on 18S rRNA sequences (Tashyreva et al. 2022), was corroborated using transcriptomic data (**Figures 43C**). These data are consistent with the evolutionary position of *S. specki* between Hemistasiidae and Diplonemidae species.

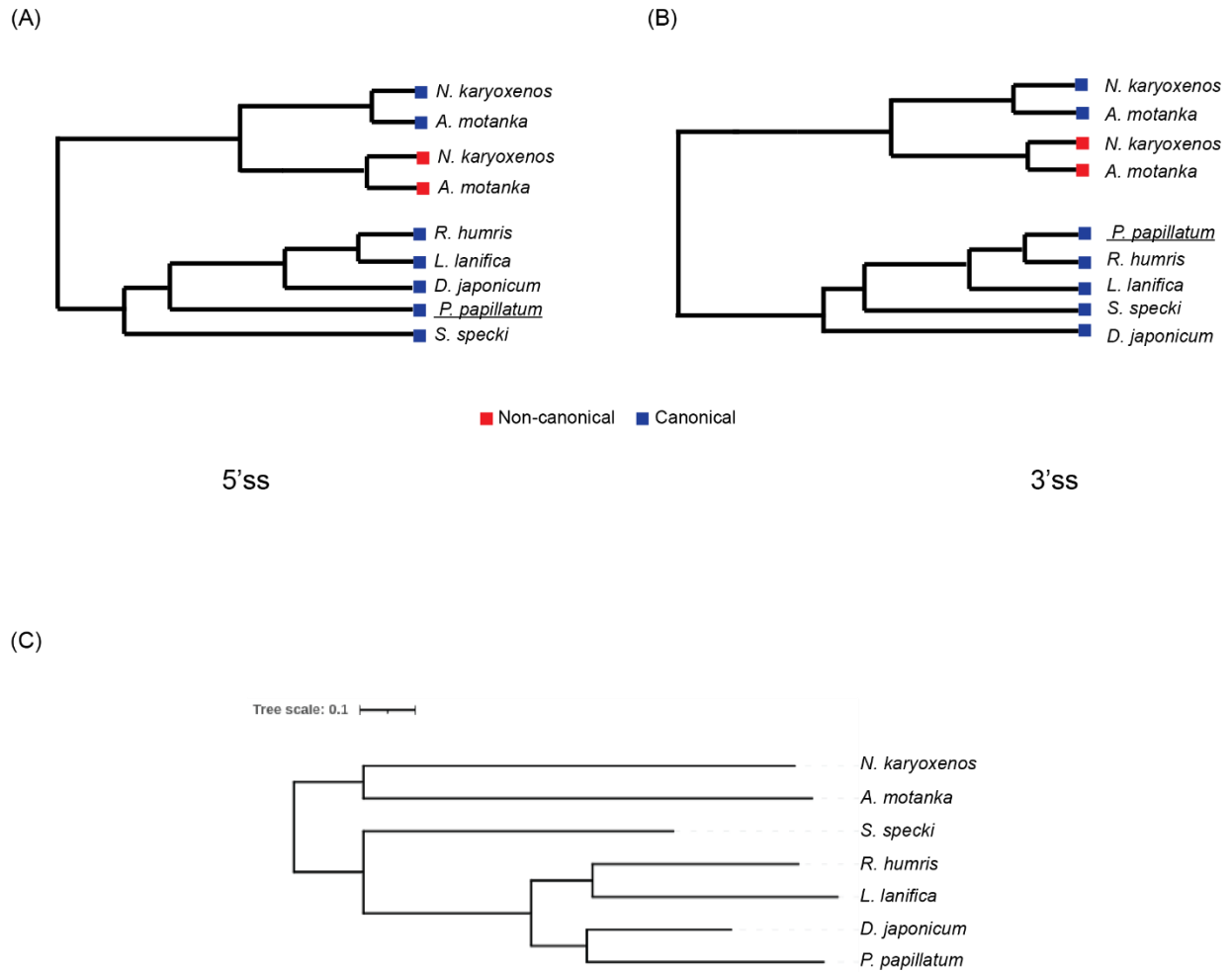


Figure 43. Clustering of diplonemids based on their splice site sequences. Phylogenetic clustering of 5' (panel A) and 3' splice site (B) motifs of canonical and non-canonical introns from seven diplonemid species. Position Weight Matrices (PWMs) were utilized to construct phylogenetic trees for the splice sites. (C) Phylogenomic tree of seven diplonemid species (*A. motanka*, *N. karyoxenos*, *S. specki*, *L. lanifica*, *D. japonicum*, *R. humris*, *P. papillatum*). Scale bar represents 0.1 substitutions per site.

8. Discussion

Roles of the 5' splice site, Polypyrimidine tract, and splicing factors in lincRNA splicing

In last ten years, lincRNAs have been the focus of intensive research, revealing that they are spliced and polyadenylated less efficiently compared to protein-coding genes (PCGs) (Seidl et al. 2006; Tilgner et al. 2012b; Mukherjee et al. 2016; Lagarde et al. 2017; Mele et al. 2017; Schlackow et al. 2017). However, the reasons for this inefficiency remain unclear. To explore this, we calculated splicing indices (SIs) for lincRNAs and PCGs expressed in human cell lines using available ENCODE RNA-Seq data (Sloan et al. 2016). Our analysis was consistent with previous findings and showed that lincRNAs are spliced less efficiently than PCGs in all the cell lines we studied (**Figure 15**). These computational results were confirmed through experimental RT-PCR analyses of several lincRNAs in our study (Krchnáková et al. 2019).

To investigate the factors influencing lincRNA splicing, we observed that lincRNAs generally have longer exons and introns compared to PCGs, which may partially account for their reduced splicing efficiency (**Figure 22A**). This observation aligns with previous studies showing that intron length is associated with splicing efficiency (Klinz and Gallwitz 1985; Sterner et al. 1996; Bell et al. 1998; Fox-Walsh et al. 2005; Dewey et al. 2006; Louloup et al. 2018). Additionally, we explored features such as secondary structure and the presence of potential splicing silencers to understand their contribution to the observed differences in splicing efficiency. However, we did not identify any significant differences between lincRNAs and PCGs in these aspects (**Figures 22 B-C, 23,24**).

Further analysis of the sequences of the 5'ss and the PPT revealed a positive correlation between the strength of these elements and splicing efficiency (**Figures 25 and 19**). Experimentally, we found that enhancing the strength of 5' and 3' splice sites significantly improved the splicing of model lincRNAs (Krchnáková et al. 2019). Our data suggest that while 5'ss and PPT are essential for splicing in general, lincRNAs require stronger splice sites, especially those with a high number of thymidine residues in the PPT, to be efficiently spliced (**Figures 16-20**). To understand why lincRNAs are more reliant on these basic splice site sequences, we examined SR protein-binding motifs, as SR proteins help facilitate splicing (reviewed in Graveley 2000; Long and Caceres 2009). We found that these motifs are more abundant in PCG exons, with

only one of 28 motifs being enriched in lincRNAs (**Figures 25A, 26, 27**). Similarly, SR proteins showed a clear binding preference for PCGs, even when comparing expression-matched lincRNAs and PCGs (**Figure 25D**). Based on these observations, we propose that lincRNAs lack the cooperative network of splicing enhancers typically present in PCGs. As a result, the splicing machinery in lincRNAs must rely heavily on the intrinsic strength of splice sites. For most lincRNAs, U1 and U2 snRNPs and their auxiliary factors must identify splice sites without the aid of splicing enhancers, making the sequences near exon-intron boundaries particularly important. Despite this, the majority of purifying selection acting on human lincRNAs is splicing-related, suggesting that the presence of introns or their splicing may serve an important, yet undiscovered, role in some lincRNAs (Schüler et al. 2014).

From an evolutionary perspective, newly emerged lincRNAs may gradually acquire functional splice sites, potentially altering their splicing and functional output over time. Cryptic 5'ss are more prevalent than cryptic 3'ss, likely due to the simpler sequence requirements for 5'ss recognition (Nakai and Sakamoto 1994). For instance, endogenous retrotransposons can provide functional 5'ss (Franke et al. 2017). Cryptic 3'ss, on the other hand, often arise from mutations that create AG dinucleotides in PPTs or enhance branch-point sequences and auxiliary signals (Královicová et al. 2005). We observed the strongest correlation between lincRNA splicing and PPT sequence, which suggests that in the absence of functional splicing enhancers, weak PPT sequence and inefficient U2AF binding represent the major barriers that evolutionary new transcripts have to overcome to become efficiently spliced.

Diplonemid introns: diverse, non-canonical, and evolutionarily intriguing

Recent research has revealed that diplonemid flagellates, previously overlooked, are not only ecologically significant but also fascinating from an evolutionary standpoint. Studies on their molecular biology, including RNA editing, mitochondrial genome organization, kinetochores, ribosomes, and trafficking machinery, have shown that diplonemids exhibit "extreme" biology. These features have evolved into highly specialized forms distinct from their textbook counterparts (Burger and Valach 2018; Lukeš et al. 2018; Kaur et al. 2020; Valach et al. 2023a; Záhonová et al. in review). The genetic structure of diplonemids also appears unconventional. Limited genomic and transcriptomic data indicate rare RNA processing events, including non-canonical introns lacking typical 5' and 3' splice site sequences. While most introns in diplonemid nuclear genes are

canonical (GT-AG), a significant fraction is non-canonical, similar to those found in euglenids (Milanowski et al. 2014; Valach et al. 2023b). Moreover, preliminary analyses of single-cell sequencing data of the most diverse and abundant diplonemid lineage – the eupelagonemids, sampled in the northern Pacific indicated a prevalence of the non-canonical introns (Gawryluk et al. 2016).

In this study, we analyzed the draft genomes and transcriptomes of six diplonemid species to identify intron types. Canonical introns dominate in four species of the Diplonemidae family, while non-canonical introns predominate in two members of the Hemistasiidae family, a sister clade to eupelagonemids. This suggests that the prevalence of non-canonical introns is a phylogenetic feature of these two clades. Interestingly, despite the differences in intron types, all six species possess complete and functional splicing machinery. A notable finding was the presence of an alternative U1 snRNA in *S. specki*, containing a C-rich region at its 5' end, potentially enabling it to pair with G-rich 5' splice sites. This adaptation may facilitate the recognition of degenerate splice sites (**Figures 30 and 39**). However, no alternative U1 snRNAs were found in *N. karyoxenos* and *A. motanka*, suggesting that in these species, the spliceosome—particularly the U1 snRNP—is not involved in splicing non-canonical introns. Non-canonical introns in *N. karyoxenos* and *A. motanka* exhibit a high potential for forming secondary structures, supported by strong base pairing between their 5' and 3' ends (**Figures 41-42**). This structural feature, consistent with earlier predictions (Milanowski et al. 2014; Gawryluk et al. 2016; Milanowski et al. 2016), may facilitate their removal.

Studies on *Euglena gracilis*, a model euglenid, showed that non-canonical introns are spliced differently from canonical ones. After excision, they form circular RNA, suggesting a spliceosome-independent pathway (Guminska et al. 2021). We propose that in diplonemids, RNA duplexes in non-canonical introns may be recognized and cleaved by RNase III enzymes, followed by exon ligation by RNA ligases, resulting in continuous mRNA and circular RNA molecules. Alternatively, a specialized machinery for non-canonical intron splicing may have evolved, as seen in tRNA intron removal pathways. This idea is supported by findings in *Drosophila melanogaster*, where tRNA introns form circular RNA after excision (Lu et al. 2015; Guminska et al. 2021). It is also possible that the classical spliceosome can adapt to splice non-canonical introns, as reported in *F. borealis* (Henriet et al. 2019). Diplonemids exhibit remarkable splicing machinery flexibility,

as seen in the conserved spliceosome across species and the unusual PPT composition in *D. japonicum*, which contains conserved adenine residues (**Figure 30**).

The origins of non-canonical introns may lie in transposable elements, such as MITE-like transposons, which can disrupt essential genes and integrate into genomes (Kazazian 2004; Feschotte and Pritham 2007). The G- and C-rich repeats at the 5' and 3' ends of non-canonical introns in *N. karyoxenos* and *A. motanka* suggest their origins from MITE-like elements (Milanowski et al. 2014). Diplonemid gene-rich genomes, with their broad metabolic capabilities and ancestral LECA-type features, highlight their exceptional adaptability. Their genomes, packed with introns, contrast sharply with their kinetoplastid relatives, which have nearly intron-free nuclear genomes (Kostygov et al. 2024). The diversity and evolution of introns in diplomids rival those in multicellular organisms (Henriet et al. 2019), showcasing the remarkable evolutionary plasticity of these marine protists.

Integration of findings: evolutionary and mechanistic parallels in splicing adaptations

The splicing mechanisms of human lincRNAs and diplomid introns, though shaped by distinct evolutionary trajectories, highlight the remarkable adaptability of RNA processing systems. Both lineages have developed unique strategies to compensate for deviations from canonical splicing signals, demonstrating how selective pressures drive splicing mechanisms to maintain functional integrity under diverse cellular and environmental constraints.

In human lincRNAs, splicing efficiency is generally lower than in protein-coding genes (PCGs) due to the absence of SR protein-mediated enhancers. To offset this, lincRNA splicing relies on stronger 5'ss and PPTs, with thymidine-rich PPTs correlating with enhanced splicing efficiency. This suggests an evolutionary pressure favoring stronger splice site sequences to compensate for the lack of enhancer-mediated regulation. Similarly, diplomids exhibit a mix of canonical and non-canonical introns, reflecting adaptations to fluctuating marine environments where temperature, salinity, and nutrient availability impose selective pressures on splicing machinery. In species with a high prevalence of non-canonical introns, particularly in the *Hemistasiidae* family, which deviate from the conserved GT-AG rule, splicing likely relies on strong RNA secondary structures at exon-intron boundaries, possibly facilitating RNA-based cleavage and ligation mechanisms independent of the spliceosome. The high frequency of these

introns suggests their role in enhancing splicing flexibility, enabling rapid adaptation to environmental stressors and providing a survival advantage in dynamic or extreme ecological niches.

The evolution of splicing mechanisms has been significantly influenced by transposable elements (TEs). Group II introns, considered ancestors of spliceosomal introns, are ribozymes capable of self-splicing and retrotransposition. Their mobility and insertion into new genomic locations have contributed to the complexity of eukaryotic gene structures. The integration of TEs into genomes can lead to the acquisition of new introns, driving the diversification of splicing mechanisms across species (Lambowitz and Belfort 2015; Zimmerly and Semper 2015).

Recent studies by Gozashti and his colleagues (Gozashti et al. 2022) identified a class of TEs termed "Introners," which can insert into multiple genes, leading to widespread intron gain. Introners, primarily derived from DNA-based transposons, are prevalent in around 5% of eukaryotic species, particularly aquatic organisms. Their horizontal transfer may contribute to intron proliferation, as evidenced in species ranging from animals to single-cell protists (McDaniel et al. 2010; Venner et al. 2017). This raises the possibility that introners could represent a fundamental and widespread source of introns across the tree of life. Although the origin of non-canonical introns in diplomonads remains unclear, the presence of G- and C-rich repeats at the 5' and 3' ends of these introns suggests a potential link to transposon activity, perhaps analogous to MITE-like transposons. Further investigation would be required to determine whether such elements play a role in the evolution of splice site diversity observed in these species.

Overall, while lincRNAs compensate for weak splicing enhancers by enhancing splice site strength, diplomonads rely on RNA structural elements to facilitate efficient splicing despite non-canonical splice sites. Both studies reveal the evolutionary adaptability of splicing mechanisms and provide deeper insights into RNA processing in two distinct biological contexts.

9. Conclusions

In my thesis, I focused on two major topics, specifically uncovering the molecular mechanisms that contribute to the generally lower splicing efficiency of lncRNAs compared to protein-coding pre-mRNAs and characterizing the canonical and non-canonical introns within diplonemid genomes and their evolutionary significance of non-canonical introns in diplonemids.

In first project, our study demonstrated that lincRNAs are spliced less efficiently than PCGs. This inefficiency is due to weaker splice site elements and the absence of a cooperative network of splicing enhancers, such as SR protein-binding motifs, that are abundant in PCGs. Consequently, lincRNA splicing depends heavily on the intrinsic strength of 5'ss and the PPT. Evolutionarily, our findings suggest that newly emerged lincRNAs gradually acquire functional splice sites, primarily through cryptic sites enabled by sequence mutations or retrotransposon insertions.

In the second project, we explored the remarkable diversity of intron composition in diplonemids, particularly the prevalence of non-canonical introns in certain clades like Hemistasiidae. Despite this diversity, all analyzed diplonemid species retain a conserved and functional splicing apparatus. In the hemistasiid diplonemids *Artemidia motanka* and *Namystynia karyoxenos*, most introns are non-canonical and lack the conserved GT-AG terminal dinucleotides typical of U2-type introns. These non-canonical introns exhibit extensive base-pairing potential, which likely brings intron ends into close proximity to facilitate splicing. Thus, while the splicing machinery is conserved across diplonemids, splice site configurations display substantial variability among species.

References:

- Abelson J, Trotta CR, Li H. 1998. tRNA splicing. *J Biol Chem* **273**: 12685-12688.
- Abou Alezz M, Celli L, Belotti G, Lisa A, Bione S. 2020. GC-AG Introns Features in Long Non-coding and Protein-Coding Genes Suggest Their Role in Gene Expression Regulation. *Front Genet* **11**.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.
- Baldi MI, Mattoccia E, Bufardecì E, Fabbri S, Tocchini-Valentini GP. 1992. Participation of the intron in the reaction catalyzed by the *Xenopus* tRNA splicing endonuclease. *Science* **255**: 1404-1408.
- Basu MK, Rogozin IB, Koonin EV. 2008. Primordial spliceosomal introns were probably U2-type. *Trends Genet* **24**: 525-528.
- Bell MV, Cowper AE, Lefranc M-P, Bell JI, Screaton GR. 1998. Influence of Intron Length on Alternative Splicing of CD44. *Mol Cell Biol* **18**: 5930-5941.
- Berriman M Ghedin E Hertz-Fowler C Blandin G Renauld H Bartholomeu DC Lennard NJ Caler E Hamlin NE Haas B et al. 2005. The genome of the African trypanosome. *Science* **309**: 416-422.
- Biniszkiwicz D, Cesnaviciene E, Shub DA. 1994. Self-splicing group I intron in cyanobacterial initiator methionine tRNA: evidence for lateral transfer of introns in bacteria. *EMBO J* **13**: 4629-4635.
- Birgisdottir ÅB, Johansen S. 2005. Site-specific reverse splicing of a HEG-containing group I intron in ribosomal RNA. *Nucleic Acids Research* **33**: 2042-2051.
- Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291-336.
- Bonen L. 1993. Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB J* **7**: 40-46.
- Botti V, McNicoll F, Steiner MC, Richter FM, Solovyeva A, Wegener M, Schwich OD, Poser I, Zarnack K, Wittig I et al. 2017. Cellular differentiation state modulates the mRNA export activity of SR proteins. *J Cell Biol* **216**: 1993-2009.
- Boulanger SC, Belcher SM, Schmidt U, Dib-Hajj SD, Schmidt T, Perlman PS. 1995. Studies of point mutants define three essential paired nucleotides in the domain 5 substructure of a group II intron. *Mol Cell Biol* **15**: 4479-4488.
- Bouvrette LPB, Cody NAL, Bergalet J, Lefebvre FA, Diot C, Wang XF, Blanchette M, Lécuyer E. 2018. CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in and human cells. *Rna* **24**: 98-113.
- Breckenridge DG, Watanabe Y, Greenwood SJ, Gray MW, Schnare MN. 1999. U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. *Proc Natl Acad Sci USA* **96**: 852-856.
- Bresson SM, Hunter OV, Hunter AC, Conrad NK. 2015. Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. *Plos Genetics* **11**.
- Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. 2002. Alternative splicing and genome complexity. *Nat Genet* **30**: 29-30.
- Bridges MC, Daulagala AC, Kourtidis A. 2021. LNCcation: lncRNA localization and function. *J Cell Biol* **220**.
- Bufardecì E, Fabbri S, Baldi MI, Mattoccia E, Tocchini-Valentini GP. 1993. In vitro genetic analysis of the structural features of the pre-tRNA required for determination of the 3' splice site in the intron excision reaction. *Embo J* **12**: 4697-4704.

- Buratti E, Stuani C, De Prato G, Baralle FE. 2007. SR protein-mediated inhibition of CFTR exon 9 inclusion: molecular characterization of the intronic splicing silencer. *Nucleic Acids Res* **35**: 4359-4368.
- Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary fates and origins of U12-type introns. *Mol Cell* **2**: 773-785.
- Burger G, Valach M. 2018. Perfection of eccentricity: Mitochondrial genomes of diplomids. *IUBMB Life* **70**: 1197-1206.
- Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. maSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**.
- Bushnell B, Rood J, Singer E. 2017. BBMerge - Accurate paired shotgun read merging via overlap. *PLoS One* **12**: e0185056.
- Cabili MN, Dunagin MC, McClanahan PD, Bialesch A, Padovan-Merhar O, Regev A, Rinn JL, Raj A. 2015. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol* **16**: 20.
- Canaday J, Tessier LH, Imbault P, Paulus F. 2001. Analysis of *Euglena gracilis* alpha-, beta- and gamma-tubulin genes: introns and pre-mRNA maturation. *Mol Genet Genomics* **265**: 153-160.
- Carlevaro-Fita J, Rahim A, Guigó R, Vardy LA, Johnson R. 2016. Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *Rna* **22**: 867-882.
- Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C et al. 2012. Long non-coding antisense RNA control translation through an embedded SINEB2 repeat. *Nature* **491**: 454-+.
- Cech TR. 1986. The Generality of Self-Splicing Rna - Relationship to Nuclear Messenger-Rna Splicing. *Cell* **44**: 207-210.
- Cech TR. 1990. Self-Splicing of Group-I Introns. *Annual Review of Biochemistry* **59**: 543-568.
- Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A, Bozzoni I. 2011. A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA (vol 147, pg 358, 2011). *Cell* **147**: 947-947.
- Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, Dinger ME, Mattick JS. 2012. Genome-wide analysis of long noncoding RNA stability. *Genome Res* **22**: 885-898.
- Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB. 2009. An Architectural Role for a Nuclear Noncoding RNA: RNA Is Essential for the Structure of Paraspeckles. *Mol Cell* **33**: 717-726.
- Clemson CM, McNeil JA, Willard HF, Lawrence JB. 1996. XIST RNA paints the inactive X chromosome at interphase: Evidence for a novel RNA involved in nuclear chromosome structure. *J Cell Biol* **132**: 259-275.
- Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Molecular Biology and Evolution* **22**: 1053-1066.
- Costa M, Michel F, Westhof E. 2000. A three-dimensional perspective on exon binding by a group II self-splicing intron. *Embo J* **19**: 5007-5018.
- Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**: 1188-1190.
- Curcio MJ, Belfort M. 1996. Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. *Cell* **84**: 9-12.

- de Lencastre A, Pyle AM. 2008. Three essential and conserved regions of the group II intron are proximal to the 5'-splice site. *RNA* **14**: 11-24.
- de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, Lara E, Berney C, Le Bescot N, Probert I et al. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.
- Decatur WA, Einvik C, Johansen S, Vogt VM. 1995. Two group I ribozymes with different functions in a nuclear rDNA intron. *Embo J* **14**: 4558-4568.
- Deveson IW, Brunck ME, Blackburn J, Tseng E, Hon T, Clark TA, Clark MB, Crawford J, Dinger ME, Nielsen LK et al. 2018. Universal Alternative Splicing of Noncoding Exons. *Cell Syst* **6**: 245-255 e245.
- Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *Bmc Genomics* **7**: 311.
- Dietrich RC, Incorvaia R, Padgett RA. 1997. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol Cell* **1**: 151-160.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.
- Dooijes D, Chaves I, Kieft R, Dirks-Mulder A, Martin W, Borst P. 2000. Base J originally found in kinetoplastida is also a minor constituent of nuclear DNA of. *Nucleic Acids Research* **28**: 3017-3021.
- Doolittle WF. 1991. The origins of introns. *Curr Biol* **1**: 145-146.
- Doolittle WF. 2013. The spliceosomal catalytic core arose in the RNA world ... or did it? *Genome Biology* **14**.
- Du Z, Sun T, Hacisuleyman E, Fei T, Wang XD, Brown M, Rinn JL, Lee MGS, Chen YW, Kantoff PW et al. 2016. Integrative analyses reveal a long noncoding RNA-mediated sponge regulatory network in prostate cancer. *Nat Commun* **7**.
- Dujon B. 1989. Group-I Introns as Mobile Genetic Elements - Facts and Mechanistic Speculations - a Review. *Gene* **82**: 91-&.
- Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L. 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**: 313-319.
- Ebel C, Frantz C, Paulus F, Imbault P. 1999. Trans-splicing and cis-splicing in the colourless Euglenoid, *Entosiphon sulcatum*. *Curr Genet* **35**: 542-550.
- Ebenezer TE, Carrington M, Lebert M, Kelly S, Field MC. 2017. Euglena gracilis Genome and Transcriptome: Organelles, Nuclear Genome Assembly Strategies and Initial Features. *Adv Exp Med Biol* **979**: 125-140.
- Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novak Vanclova AMG, Prasad B, Soukal P, Santana-Molina C, O'Neill E, Nankissoor NN et al. 2019. Transcriptome, proteome and draft genome of *Euglena gracilis*. *BMC Biol* **17**: 11.
- Edgell DR. 2009. Selfish DNA: Homing Endonucleases Find a Home. *Current Biology* **19**: R115-R117.
- Emms DM, Kelly S. 2017. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Mol Biol Evol* **34**: 3267-3278.
- Emms DM, Kelly S. 2018. STAG: Species Tree Inference from All Genes *bioRxiv*.
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**: 238.

- Englert M, Sheppard K, Aslanian A, Yates JR, 3rd, Soll D. 2011. Archaeal 3'-phosphate RNA splicing ligase characterization identifies the missing component in tRNA maturation. *Proc Natl Acad Sci U S A* **108**: 1290-1295.
- Erkelenz S, Mueller WF, Evans MS, Busch A, Schöneweis K, Hertel KJ, Schaal H. 2013. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA* **19**: 96-102.
- Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single Nucleotide Polymorphism-Based Validation of Exonic Splicing Enhancers. *PLoS Biol* **2**: e268.
- Fazal FM, Han S, Parker KR, Kaewsapsak P, Xu J, Boettiger AN, Chang HY, Ting AY. 2019. Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell* **178**: 473-+.
- Fedorova O, Mitros T, Pyle AM. 2003. Domains 2 and 3 interact to form critical elements of the group II intron active site. *J Mol Biol* **330**: 197-209.
- Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41**: 331-368.
- Fica SM, Tuttle N, Novak T, Li NS, Lu J, Koodathingal P, Dai Q, Staley JP, Piccirilli JA. 2013. RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**: 229-+.
- Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, Bowler C, Lukes J, Horak A. 2016. Extreme Diversity of Diplonemid Eukaryotes in the Ocean. *Curr Biol* **26**: 3060-3065.
- Fox-Walsh KL, Dou Y, Lam BJ, Hung S-p, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci USA* **102**: 16176-16181.
- Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, Kuzman M, Fulka H, Cernohorska M, Urbanova J et al. 2017. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Res* **27**: 1384-1394.
- Fujishima K, Sugahara J, Miller CS, Baker BJ, Di Giulio M, Takesue K, Sato A, Tomita M, Banfield JF, Kanai A. 2011. A novel three-unit tRNA splicing endonuclease found in ultrasmall Archaea possesses broad substrate specificity. *Nucleic Acids Research* **39**: 9695-9704.
- Gallego ME, Gattoni R, Stévenin J, Marie J, Expert-Bezançon A. 1997. The SR splicing factors ASF/SF2 and SC35 have antagonistic effects on intronic enhancer-dependent splicing of the β -tropomyosin alternative exon 6A. *EMBO J* **16**: 1772-1784.
- Garg K, Green P. 2007. Differing patterns of selection in alternative and constitutive splice sites. *Genome Res* **17**: 1015-1022.
- Gawryluk RMR, Del Campo J, Okamoto N, Strassert JFH, Lukes J, Richards TA, Worden AZ, Santoro AE, Keeling PJ. 2016. Morphological Identification and Single-Cell Genomics of Marine Diplonemids. *Curr Biol* **26**: 3053-3059.
- Gehring NH, Roignant JY. 2021. Anything but Ordinary - Emerging Splicing Mechanisms in Eukaryotic Gene Regulation. *Trends Genet* **37**: 355-372.
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.
- Gerber JL, Kohler S, Peschek J. 2022. Eukaryotic tRNA splicing - one goal, two strategies, many players. *Biol Chem* **403**: 765-778.
- Gilbert W. 1987. The exon theory of genes. *Cold Spring Harb Symp Quant Biol* **52**: 901-905.
- Gilbert W, Glynias M. 1993. On the ancient nature of introns. *Gene* **135**: 137-144.

- Gong CG, Maquat LE. 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**: 284-+.
- Gozashti L, Roy SW, Thornlow B, Kramer A, Ares M, Jr., Corbett-Detig R. 2022. Transposable elements drive intron gain in diverse eukaryotes. *Proc Natl Acad Sci U S A* **119**: e2209766119.
- Graveley BR. 2000. Sorting out the complexity of SR protein functions. *RNA* **6**: 1197-1211.
- Graveley BR, Kaur A, Gunning D, Zipursky SL, Rowen L, Clemens JC. 2004. The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes. *Rna* **10**: 1499-1506.
- Grelet S, Link LA, Howley B, Obellianne C, Palanisamy V, Gangaraju VK, Diehl JA, Howe PH. 2017. A regulated mRNA to lncRNA splice switch mediates EMT and tumour progression (vol 19, pg 1105, 2017). *Nat Cell Biol* **19**: 1443-1443.
- Guminska N, Plecha M, Zakrys B, Milanowski R. 2018. Order of removal of conventional and nonconventional introns from nuclear transcripts of *Euglena gracilis*. *PLoS Genet* **14**: e1007761.
- Guminska N, Zakrys B, Milanowski R. 2021. A New Type of Circular RNA derived from Nonconventional Introns in Nuclear Genes of Euglenids. *J Mol Biol* **433**: 166758.
- Guo CJ, Ma XK, Xing YH, Zheng CC, Xu YF, Shan L, Zhang J, Wang S, Wang Y, Carmichael GG et al. 2020. Distinct Processing of lncRNAs Contributes to Non-conserved Functions in Stem Cells. *Cell* **181**: 621-636 e622.
- Haberman N, Huppertz I, Attig J, König J, Wang Z, Hauer C, Hentze MW, Kulozik AE, Le Hir H, Curk T et al. 2017. Insights into the design and interpretation of iCLIP experiments. *Genome Biol* **18**: 7.
- Haerty W, Ponting CP. 2015. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* **21**: 333-346.
- Hall SL, Padgett RA. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J Mol Biol* **239**: 357-365.
- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760-1774.
- Haugen P, Andreassen M, Birgisdottir AB, Johansen S. 2004. Hydrolytic cleavage by a group I intron ribozyme is dependent on RNA structures not important for splicing. *Eur J Biochem* **271**: 1015-1024.
- Haugen P, Simon DM, Bhattacharya D. 2005. The natural history of group I introns. *Trends Genet* **21**: 111-119.
- Henriet S, Colom Sanmarti B, Sumic S, Chourrout D. 2019. Evolution of the U2 Spliceosome for Processing Numerous and Highly Diverse Non-canonical Introns in the Chordate *Fritillaria borealis*. *Curr Biol* **29**: 3193-3199.
- Henze K, Badr A, Wettern M, Cerff R, Martin W. 1995. A Nuclear Gene of Eubacterial Origin in *Euglena-Gracilis* Reflects Cryptic Endosymbioses during Protist Evolution. *Proc Natl Acad Sci USA* **92**: 9122-9126.
- Hirata A, Fujishima K, Yamagami R, Kawamura T, Banfield JF, Kanai A, Hori H. 2012. X-ray structure of the fourth type of archaeal tRNA splicing endonuclease: insights into the evolution of a novel three-unit composition and a unique loop involved in broad substrate specificity. *Nucleic Acids Research* **40**: 10554-10566.

- Hoskins AA, Friedman LJ, Gallagher SS, Crawford DJ, Anderson EG, Wombacher R, Ramirez N, Cornish VW, Gelles J, Moore MJ. 2011. Ordered and dynamic assembly of single spliceosomes. *Science* **331**: 1289-1295.
- Hoskins AA, Moore MJ. 2012. The spliceosome: a flexible, reversible macromolecular machine. *Trends in biochemical sciences* **37**: 179-188.
- Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, Chess A. 2007. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *Bmc Genomics* **8**.
- Chakrabarti AM, Haberman N, Praznik A, Luscombe NM, Ule J. 2018. Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies. *Annual Review of Biomedical Data Science* **1**: 235-261.
- Irimia M, Roy SW. 2014. Origin of Spliceosomal Introns and Alternative Splicing. *CSH Perspect Biol* **6**.
- Ivens AC Peacock CS Worthey EA Murphy L Aggarwal G Berriman M Sisk E Rajandream MA Adlem E Aert R et al. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**: 436-442.
- Iwata H, Gotoh O. 2011. Comparative analysis of information contents relevant to recognition of introns in many species. *Bmc Genomics* **12**.
- Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, Hartley C, Sanders M, Wastling JM, Dacks JB et al. 2016. Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. *Curr Biol* **26**: 161-172.
- Jackson IJ. 1991. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res* **19**: 3795-3798.
- Jacquier A, Michel F. 1987. Multiple exon-binding sites in class II self-splicing introns. *Cell* **50**: 17-29.
- Jiang Z-h, Zhang W-j, Rao Y, Wu JY. 1998. Regulation of Ich-1 pre-mRNA alternative splicing and apoptosis by mammalian splicing factors. *Proc Natl Acad Sci USA* **95**: 9155-9160.
- Johansen S, Einvik C, Nielsen H. 2002. DiGIR1 and NaGIR1: naturally occurring group I-like ribozymes with unique core organization and evolved biological role. *Biochimie* **84**: 905-912.
- Johansen S, Vogt VM. 1994. An intron in the nuclear ribosomal DNA of *Didymium iridis* codes for a group I ribozyme and a novel ribozyme that cooperate in self-splicing. *Cell* **76**: 725-734.
- Kaewsapsak P, Shechner DM, Mallard W, Rinn JL, Ting AY. 2017. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *Elife* **6**.
- Kanopka A, Mühlemann O, Akusjärvi G. 1996. Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature* **381**: 535.
- Kaur B, Valach M, Pena-Diaz P, Moreira S, Keeling PJ, Burger G, Lukes J, Faktorova D. 2018. Transformation of *Diplonema papillatum*, the type species of the highly diverse and abundant marine microeukaryotes Diplonemida (Euglenozoa). *Environ Microbiol* **20**: 1030-1040.
- Kaur B, Záhonová K, Valach M, Faktorová D, Prokopchuk G, Burger G, Lukes J. 2020. Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. *Nucleic Acids Res* **48**: 2694-2708.
- Kazazian HH, Jr. 2004. Mobile elements: drivers of genome evolution. *Science* **303**: 1626-1632.

- Keating KS, Toor N, Perlman PS, Pyle AM. 2010. A structural analysis of the group II intron active site and implications for the spliceosome. *RNA* **16**: 1-9.
- Kennell JC, Moran JV, Perlman PS, Butow RA, Lambowitz AM. 1993. Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. *Cell* **73**: 133-146.
- Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews Genetics* **11**: 345-355.
- Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Morales DR, Thomas K, Presser A, Bernstein BE, van Oudenaarden A et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *P Natl Acad Sci USA* **106**: 11667-11672.
- Khan MR, Wellinger RJ, Laurent B. 2021. Exploring the Alternative Splicing of Long Noncoding RNAs. *Trends Genet* **37**: 695-698.
- Klinz F-J, Gallwitz D. 1985. Size and position of intervening sequences are critical for the splicing efficiency of pre-mRNA in the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res* **13**: 3791-3804.
- König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909.
- Konkel DA, Tilghman SM, Leder P. 1978. Sequence of the Chromosomal Mouse Beta-Globin Major Gene - Homologies in Capping, Splicing and Poly(a) Sites. *Cell* **15**: 1125-1132.
- Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* **1**: 22.
- Kostygov AY, Skypalova K, Kraeva N, Kalita E, McLeod C, Yurchenko V, Field MC, Lukes J, Butenko A. 2024. Comprehensive analysis of the Kinetoplastea intron landscape reveals a novel intron-containing gene and the first exclusively trans-splicing eukaryote. *BMC Biol* **22**: 281.
- Kráľovicová J, Christensen MB, Vorechovský I. 2005. Biased exon/intron distribution of cryptic and de novo 3' splice sites. *Nucleic Acids Res* **33**: 4882-4898.
- Krchnáková Z, Thakur PK, Krausová M, Bieberstein N, Haberman N, Müller-McNicoll M, Stanek D. 2019. Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5 splice-site sequences due to weak interactions with SR proteins. *Nucleic Acids Res* **47**: 911-928.
- Kugel JF, Goodrich JA. 2012. Non-coding RNAs: key regulators of mammalian transcription. *Trends Biochem Sci* **37**: 144-151.
- Kuhse MG, Strickland R, Palmer JD. 1990. An ancient group I intron shared by eubacteria and chloroplasts. *Science* **250**: 1570-1573.
- Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* **49**: 1731.
- Lambowitz AM, Belfort M. 2015. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* **3**.
- Lara E, Moreira D, Vereshchaka A, Lopez-Garcia P. 2009. Pan-oceanic distribution of new highly diverse clades of deep-sea diplomonads. *Environ Microbiol* **11**: 47-55.
- Latgé G, Poulet C, Bours V, Josse C, Jerusalem G. 2018. Natural Antisense Transcripts: Molecular Mechanisms and Implications in Breast Cancers. *Int J Mol Sci* **19**.

- Lawrence M, Gentleman R, Carey V. 2009. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**: 1841-1842.
- Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comp Biol* **9**: e1003118.
- Lee S, Kopp F, Chang TC, Sataluri A, Chen BB, Sivakumar S, Yu HT, Xie Y, Mendell JT. 2016. Noncoding RNA Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164**: 69-80.
- Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**: W293-W296.
- Li H, Abelson J. 2000. Crystal structure of a dimeric archaeal splicing endonuclease. *J Mol Biol* **302**: 639-648.
- Li H, Trotta CR, Abelson J. 1998. Crystal structure and evolution of a transfer RNA splicing enzyme. *Science* **280**: 279-284.
- Li P, Tao Z, Dean C. 2015. Phenotypic evolution through variation in splicing of the noncoding RNA COOLAIR. *Genes Dev* **29**: 696-701.
- Liang XH, Haritan A, Uliel S, Michaeli S. 2003. trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot Cell* **2**: 830-840.
- Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, Chaffron S, Ignacio-Espinosa JC, Roux S, Vincent F et al. 2015. Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* **348**: 1262073.
- Lin AF, Li CL, Xing Z, Hu QS, Liang K, Han L, Wang C, Hawke DH, Wang SY, Zhang YY et al. 2016. The -lncRNA activates normoxic HIF1 α signalling in triple-negative breast cancer. *Nat Cell Biol* **18**: 213-+.
- Lin CF, Mount SM, Jarmolowski A, Makalowski W. 2010. Evolutionary dynamics of U12-type spliceosomal introns. *BMC Evol Biol* **10**.
- Logsdon JM, Jr. 1998. The recent origins of spliceosomal introns revisited. *Curr Opin Genet Dev* **8**: 637-648.
- Long Jennifer C, Caceres Javier F. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**: 15-27.
- Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C, Moreira D. 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**: 603-607.
- Lorenz R, Bernhart SH, Siederdisen CHZ, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithm Mol Biol* **6**.
- Louloupi A, Ntini E, Conrad T, Ørom UAV. 2018. Transient N-6-Methyladenosine Transcriptome Sequencing Reveals a Regulatory Role of m6A in Splicing Efficiency. *Cell Rep* **23**: 3429-3437.
- Lu ZP, Filonov GS, Noto JJ, Schmidt CA, Hatkevich TL, Wen Y, Jaffrey SR, Matera AG. 2015. Metazoan tRNA introns generate stable circular RNAs in vivo. *RNA* **21**: 1554-1565.
- Lukes J, Flegontova O, Horak A. 2015. Diplonemids. *Curr Biol* **25**: R702-704.
- Lukeš J, Wheeler R, Jirsova D, David V, Archibald JM. 2018. Massive mitochondrial DNA content in diplonemid and kinetoplastid protists. *IUBMB Life* **70**: 1267-1274.
- Lynch M, Richardson AO. 2002. The evolution of spliceosomal introns. *Curr Opin Genet Dev* **12**: 701-710.
- Ma LN, Bajic VB, Zhang Z. 2013. On the classification of long non-coding RNAs. *Rna Biol* **10**: 925-934.

- Martin W, Koonin EV. 2006a. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440**: 41-45.
- Martin W, Koonin EV. 2006b. A positive definition of prokaryotes. *Nature* **442**: 868-868.
- Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, Chen R, Dean C, Dinger ME, Fitzgerald KA et al. 2023. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* **24**: 430-447.
- McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. 2010. High frequency of horizontal gene transfer in the oceans. *Science* **330**: 50.
- McWatters DC, Russell AG. 2017. Euglena Transcript Processing. *Adv Exp Med Biol* **979**: 141-158.
- Mele M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. 2017. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res* **27**: 27-37.
- Melé M, Rinn JL. 2016. "Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription. *Mol Cell* **62**: 657-664.
- Miao H, Wang LL, Zhan HM, Dai JS, Chang YB, Wu F, Liu T, Liu ZY, Gao CF, Li L et al. 2019. A long noncoding RNA distributed in both nucleus and cytoplasm operates in the PYCARD-regulated apoptosis by coordinating the epigenetic and translational regulation. *Plos Genetics* **15**.
- Michel F, Umesono K, Ozeki H. 1989. Comparative and functional anatomy of group II catalytic introns--a review. *Gene* **82**: 5-30.
- Michel F, Westhof E. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol* **216**: 585-610.
- Mikheeva S, Murray HL, Zhou H, Turczyk BM, Jarrell KA. 2000. Deletion of a conserved dinucleotide inhibits the second step of group II intron splicing. *RNA* **6**: 1509-1515.
- Milanowski R, Guminska N, Karnkowska A, Ishikawa T, Zakrys B. 2016. Intermediate introns in nuclear genes of euglenids - are they a distinct type? *BMC Ecol Evo* **16**.
- Milanowski R, Karnkowska A, Ishikawa T, Zakrys B. 2014. Distribution of conventional and nonconventional introns in tubulin (alpha and beta) genes of euglenids. *Mol Biol Evol* **31**: 584-593.
- Min XJ, Butler G, Storms R, Tsang A. 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res* **33**: W677-680.
- Missler M, Sudhof TC. 1998. Neurexins: three genes and 1001 products. *Trends in genetics : TIG* **14**: 20-26.
- Mondal T, Rasmussen M, Pandey GK, Isaksson A, Kanduri C. 2010. Characterization of the RNA content of chromatin. *Genome Res* **20**: 899-907.
- Moyer DC, Larue GE, Hershberger CE, Roy SW, Padgett RA. 2020. Comprehensive database and evolutionary dynamics of U12-type introns. *Nucleic Acids Res* **48**: 7066-7078.
- Mueller WF, Hertel KJ. 2011. The role of SR and SR-related proteins in pre-mRNA splicing. In *RNA Binding Proteins*, Vol I (ed. Z Lorkovic), pp. 1-21. Landes Bioscience and Springer Science+Business Media, New York.
- Muchhal US, Schwartzbach SD. 1992. Characterization of a Euglena gene encoding a polypeptide precursor to the light-harvesting chlorophyll a/b-binding protein of photosystem II. *Plant Mol Biol* **18**: 287-299.

- Muchhal US, Schwartzbach SD. 1994. Characterization of the Unique Intron-Exon Junctions of Euglena Gene(S) Encoding the Polyprotein Precursor to the Light-Harvesting Chlorophyll a/B Binding-Protein of Photosystem-II. *Nucleic Acids Research* **22**: 5737-5744.
- Mukherjee N, Calviello L, Hirsekorn A, de Pretis S, Pelizzola M, Ohler U. 2016. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat Struct Mol Biol* **24**: 86.
- Müller-McNicoll M, Botti V, de Jesus Domingues AM, Brandl H, Schwich OD, Steiner MC, Curk T, Poser I, Zarnack K, Neugebauer KM. 2016. SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes Dev* **30**: 553-566.
- Mulligan GJ, Guo W, Wormsley S, Helfman DM. 1992. Polypyrimidine tract binding protein interacts with sequences involved in alternative splicing of beta-tropomyosin pre-mRNA. *J Biol Chem* **267**: 25480-25487.
- Nakai K, Sakamoto H. 1994. Construction of a novel database containing aberrant splicing mutations of mammalian genes. *Gene* **141**: 171-177.
- Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933-2935.
- Nielsen H, Fiskaa T, Birgisdottir AB, Haugen P, Einvik C, Johansen S. 2003. The ability to form full-length intron RNA circles is a general property of nuclear group I introns. *RNA* **9**: 1464-1475.
- Nielsen H, Johansen SD. 2007. A new RNA branching activity: The GIR1 ribozyme. *Blood Cell Mol Dis* **38**: 102-109.
- Nielsen H, Johansen SD.. 2009. Group I introns Moving in new directions. *Rna Biology* **6**: 375-383.
- Nielsen H, Westhof E, Johansen S. 2005. An mRNA is capped by a 2', 5' lariat catalyzed by a group I-like ribozyme. *Science* **309**: 1584-1587.
- Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457-463.
- Orphanides G, Reinberg D. 2002. A unified theory of gene expression. *Cell* **108**: 439-451.
- Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2024. Biostrings: Efficient manipulation of biological strings. R package version 2.70.3, <https://bioconductor.org/packages/Biostrings>.
- Palmer JD, Logsdon JM. 1991. The recent origins of introns. *Curr Opin Genet Dev* **1**: 470-477.
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413-1415.
- Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**: 526-528.
- Patel AA, Steitz JA. 2003. Splicing double: Insights from the second spliceosome. *Nat Rev Mol Cell Bio* **4**: 960-970.
- Paushkin SV, Patel M, Furia BS, Peltz SW, Trotta CR. 2004. Identification of a human endonuclease complex reveals a link between tRNA splicing and Pre-mRNA 3' end formation. *Cell* **117**: 311-321.
- Paz I, Akerman M, Dror I, Kosti I, Mandel-Gutfreund Y. 2010. SFmap: a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Res* **38**: W281-285.
- Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL. 1986. A self-splicing RNA excises an intron lariat. *Cell* **44**: 213-223.

- Penny D, Hoepfner MP, Poole AM, Jeffares DC. 2009. An overview of the introns-first theory. *J Mol Evol* **69**: 527-540.
- Phizicky EM, Schwartz RC, Abelson J. 1986. *Saccharomyces cerevisiae* tRNA ligase. Purification of the protein and isolation of the structural gene. *J Biol Chem* **261**: 2978-2986.
- Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* **136**: 629-641.
- Poole AM, Jeffares DC, Penny D. 1998. The path from the RNA world. *J Mol Evol* **46**: 1-17.
- Popow J, Englert M, Weitzer S, Schleiffer A, Mierzwa B, Mechtler K, Trowitzsch S, Will CL, Lührmann R, Söll D et al. 2011. HSPC117 Is the Essential Subunit of a Human tRNA Splicing Ligase Complex. *Science* **331**: 760-764.
- Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes De Novo Assembler. *Curr Protoc Bioinformatics* **70**: e102.
- Prokopchuk G, Tashyreva D, Yabuki A, Horak A, Masarova P, Lukes J. 2019. Morphological, Ultrastructural, Motility and Evolutionary Characterization of Two New Hemistasiidae Species. *Protist* **170**: 259-282.
- Převorovský M, Hálová M, Abrahámová K, Libus J, Folk P. 2016. Workflow for Genome-Wide Determination of Pre-mRNA Splicing Efficiency from Yeast RNA-seq Data. *Biomed Res Int* **2016**: 4783841.
- Pyle AM, Lambowitz AM. 2006. Group II Introns: Ribozymes That Splice RNA and Invade DNA. *Cold Spring Harb Mon* **43**: 469-505.
- Qian Q, Keeling PJ. 2001. Diplonemid glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and prokaryote-to-eukaryote lateral gene transfer. *Protist* **152**: 193-201.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.
- Rauhut R, Green PR, Abelson J. 1990. Yeast Transfer Rna-Splicing Endonuclease Is a Heterotrimeric Enzyme. *J Biol Chem* **265**: 18180-18184.
- Reinhold-Hurek B, Shub DA. 1992. Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature* **357**: 173-176.
- Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**: 129.
- Reyes VM, Abelson J. 1988. Substrate recognition and splice site determination in yeast tRNA splicing. *Cell* **55**: 719-730.
- Robart AR, Zimmerly S. 2005. Group II intron retroelements: function and diversity. *Cytogenet Genome Res* **110**: 589-597.
- Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. *Biology Direct* **7**.
- Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* **13**: 1512-1517.
- Roman J, Woodson SA. 1998. Integration of the group I intron into bacterial rRNA by reverse splicing. *P Natl Acad Sci USA* **95**: 2134-2139.
- Roy B, Haupt LM, Griffiths LR. 2013. Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity. *Curr Genomics* **14**: 182-194.
- Russell AG, Charette JM, Spencer DF, Gray MW. 2006. An early evolutionary origin for the minor spliceosome. *Nature* **443**: 863-866.

- Russell AG, Watanabe Y, Charette JM, Gray MW. 2005. Unusual features of fibrillarin cDNA and gene structure in *Euglena gracilis*: evolutionary conservation of core proteins and structural predictions for methylation-guide box C/D snoRNPs throughout the domain Eucarya. *Nucleic Acids Res* **33**: 2781-2791.
- S M Berget, C Moore, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**: 3171-3175.
- Saldanha R, Mohr G, Belfort M, Lambowitz AM. 1993. Group I and group II introns. *FASEB J* **7**: 15-24.
- Saxena A, Carninci P. 2011. Long non-coding RNA modifies chromatin Epigenetic silencing by long non-coding RNAs. *Bioessays* **33**: 830-839.
- Scamborova P, Wong A, Steitz JA. 2004. An intronic enhancer regulates splicing of the twintron of pre-mRNA by two different spliceosomes. *Molecular and Cellular Biology* **24**: 1855-1869.
- Seidl CI, Stricker SH, Barlow DP. 2006. The imprinted *Air* ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. *EMBO J* **25**: 3565-3575.
- Sharp PA. 1985. On the Origin of Rna Splicing and Introns. *Cell* **42**: 397-400.
- Sharp PA. 1991. "Five easy pieces". *Science* **254**: 663.
- Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ. 2017. Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. *Mol Cell* **65**: 25-38.
- Schmidt U, Podar M, Stahl U, Perlman PS. 1996. Mutations of the two-nucleotide bulge of D5 of a group II intron block splicing in vitro and in vivo: Phenotypes and suppressor mutations. *Rna* **2**: 1161-1172.
- Schneider C, Will CL, Makarova OV, Makarov EM, Lührmann R. 2002. Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. *Molecular and Cellular Biology* **22**: 3219-3229.
- Schüler A, Ghanbarian AT, Hurst LD. 2014. Purifying Selection on Splice-Related Motifs, Not Expression Level nor RNA Folding, Explains Nearly All Constraint on Human lincRNAs. *Mol Biol Evol* **31**: 3164-3183.
- Simard MJ, Chabot B. 2002. SRp30c Is a Repressor of 3' Splice Site Utilization. *Mol Cell Biol* **22**: 4001-4010.
- Simpson AGB. 1997. The identity and composition of the Euglenozoa. *Arch Protistenkd* **148**: 318-328.
- Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT et al. 2016. ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**: D726-732.
- Staley JP, Guthrie C. 1998. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**: 315-326.
- Sterner DA, Carlo T, Berget SM. 1996. Architectural limits on split genes. *Proc Natl Acad Sci USA* **93**: 15081-15085.
- Stoltzfus A, Logsdon JM, Jr., Palmer JD, Doolittle WF. 1997. Intron "sliding" and the diversity of intron positions. *Proc Natl Acad Sci U S A* **94**: 10739-10744.
- Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Jr., Doolittle WF. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science* **265**: 202-207.
- Sturm NR, Maslov DA, Grisard EC, Campbell DA. 2001. Diplonema spp. possess spliced leader RNA genes similar to the kinetoplastida. *J Eukaryot Microbiol* **48**: 325-331.

- Sun QY, Hao QY, Prasanth KV. 2018. Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. *Trends Genet* **34**: 142-157.
- Tarn WY, Steitz JA. 1997. Pre-mRNA splicing: The discovery of a new spliceosome doubles the challenge. *Trends Biochem Sci* **22**: 132-137.
- Tashyreva D, Prokopchuk G, Yabuki A, Kaur B, Faktorova D, Votypka J, Kusaka C, Fujikura K, Shiratori T, Ishida KI et al. 2018. Phylogeny and Morphology of New Diplonemids from Japan. *Protist* **169**: 158-179.
- Tashyreva D, Simpson AGB, Prokopchuk G, Skodová-Sveráková I, Butenko A, Hammond M, George EE, Flegontova O, Záhonová K, Faktorová D et al. 2022. Diplonemids-A Review on "New" Flagellates on the Oceanic Block. *Protist* **173**.
- Team RC. 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org>.
- ten Dam GB, Zilch CF, Wallace D, Wieringa B, Beverley PCL, Poels LG, Screaton GR. 2000. Regulation of Alternative Splicing of CD45 by Antagonistic Effects of SR Protein Splicing Factors. *J Immunol* **164**: 5287-5295.
- Tessier LH, Paulus F, Keller M, Vial C, Imbault P. 1995. Structure and expression of *Euglena gracilis* nuclear rbcS genes encoding the small subunits of the ribulose 1,5-bisphosphate carboxylase/oxygenase: a novel splicing process for unusual intervening sequences? *J Mol Biol* **245**: 22-33.
- Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* **Chapter 2**: Unit 2 3.
- Tilgner H, Knowles DG, Johnson R, Davis CA, Chakraborty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R. 2012a. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22**: 1616-1625.
- Tilgner H, Knowles DG, Johnson R, Davis CA, Chakraborty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R. 2012b. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22**: 1616-1625.
- Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP. 2005. Structure, function, and evolution of the tRNA endonucleases of Archaea: An example of subfunctionalization. *P Natl Acad Sci USA* **102**: 8933-8938.
- Toor N, Keating KS, Taylor SD, Pyle AM. 2008. Crystal structure of a self-spliced group II intron. *Science* **320**: 77-82.
- Triemer RE, Farmer MA. 1991. An Ultrastructural Comparison of the Mitotic Apparatus, Feeding Apparatus, Flagellar Apparatus and Cytoskeleton in Euglenoids and Kinetoplastids. *Protoplasma* **164**: 91-104.
- Trotta CR, Miao F, Arn EA, Stevens SW, Ho CK, Rauhut R, Abelson JN. 1997. The yeast tRNA splicing endonuclease: A tetrameric enzyme with two active site subunits homologous to the archaeal tRNA endonucleases. *Cell* **89**: 849-858.
- Trotta CR, Paushkin SV, Patel M, Li H, Peltz SW. 2006. Cleavage of pre-tRNAs by the splicing endonuclease requires a composite active site. *Nature* **441**: 375-377.
- Trowitzsch S, Weber G, Luhrmann R, Wahl MC. 2009. Crystal structure of the Pml1p subunit of the yeast precursor mRNA retention and splicing complex. *Journal of molecular biology* **385**: 531-541.

- Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science* **329**: 689-693.
- Turunen JJ, Niemela EH, Verma B, Frilander MJ. 2013. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA* **4**: 61-76.
- Ulitsky I, Bartel DP. 2013. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* **154**: 26-46.
- Valach M, Benz C, Aguilar LC, Gahura O, Faktorová D, Zíková A, Oeffinger M, Burger G, Gray MW, Lukes J. 2023a. Miniature RNAs are embedded in an exceptionally protein-rich mitoribosome via an elaborate assembly pathway. *Nucleic Acids Res* **51**: 6443-6460.
- Valach M, Moreira S, Petitjean C, Benz C, Butenko A, Flegontova O, Nenarokova A, Prokopchuk G, Batstone T, Lapebie P et al. 2023b. Recent expansion of metabolic versatility in *Diplonema papillatum*, the model species of a highly speciose group of marine eukaryotes. *BMC Biol* **21**: 99.
- van den Hoogenhof MM, Pinto YM, Creemers EE. 2016. RNA Splicing: Regulation and Dysregulation in the Heart. *Circ Res* **118**: 454-468.
- van Heesch S, van Iterson M, Jacobi J, Boymans S, Essers PB, de Bruijn E, Hao WS, MacInnes AW, Cuppen E, Simonis M. 2014. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biology* **15**.
- Venner S, Miele V, Terzian C, Biemont C, Daubin V, Feschotte C, Pontier D. 2017. Ecological networks to unravel the routes to horizontal transposon transfers. *PLoS Biol* **15**: e2001536.
- Vesteg M, Vacula R, Steiner JM, Mateasikova B, Loffelhardt W, Brejova B, Krajcovic J. 2010. A Possible Role for Short Introns in the Acquisition of Stroma-Targeting Peptides in the Flagellate *Euglena gracilis*. *DNA Res* **17**: 223-231.
- Villa T, Pleiss JA, Guthrie C. 2002. Spliceosomal snRNAs:: Mg-dependent chemistry at the catalytic core? *Cell* **109**: 149-152.
- Wagner EJ, Garcia-Blanco MA. 2001. Polypyrimidine Tract Binding Protein Antagonizes Exon Definition. *Mol Cell Biol* **21**: 3281-3288.
- Wang Y, Wang J, Gao L, Lafyatis R, Stamm S, Andreadis A. 2005. Tau Exons 2 and 10, Which Are Misregulated in Neurodegenerative Diseases, Are Partly Regulated by Silencers Which Bind a SRp30c-SRp55 Complex That Either Recruits or Antagonizes htra2 β 1. *J Biol Chem* **280**: 14230-14239.
- Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM. 1999. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell* **4**: 239-250.
- Waseem M, Liu YL, Xia R. 2021. Long Non-Coding RNAs, the Dark Matter: An Emerging Regulatory Component in Plants. *Int J Mol Sci* **22**.
- Wei W, Pelechano V, Järvelin AI, Steinmetz LM. 2011. Functional consequences of bidirectional promoters. *Trends Genet* **27**: 267-276.
- Weinberg Z, Breaker RR. 2011. R2R--software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics* **12**: 3.
- Westaway SK, Phizicky EM, Abelson J. 1988. Structure and function of the yeast tRNA ligase gene. *J Biol Chem* **263**: 3171-3176.
- Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.
- Wilkinson ME, Charenton C, Nagai K. 2020. RNA Splicing by the Spliceosome. *Annu Rev Biochem* **89**: 359-388.

- Will CL, Luhrmann R. 2005. Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol Chem* **386**: 713-724.
- Will CL, Luhrmann R. 2011. Spliceosome Structure and Function. *CSH Perspect Biol* **3**.
- Wu HJ, GaubierComella P, Delseny M, Grellet F, VanMontagu M, Rouze P. 1996. Non-canonical introns are at least 10(9) years old. *Nat Genet* **14**: 383-384.
- Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**: 1859-1875.
- Xiao X, Wang Z, Jang M, Burge CB. 2007. Coevolutionary networks of splicing cis-regulatory elements. *Proc Natl Acad Sci USA* **104**: 18583-18588.
- Xu MQ, Kathe SD, Goodrichblair H, Nierzwickibauer SA, Shub DA. 1990. Bacterial Origin of a Chloroplast Intron - Conserved Self-Splicing Group-I Introns in Cyanobacteria. *Science* **250**: 1566-1570.
- Xue S, Calvin K, Li H. 2006. RNA recognition and cleavage by a splicing endonuclease. *Science* **312**: 906-910.
- Xue Y, Zhou Y, Wu T, Zhu T, Ji X, Kwon Y-S, Zhang C, Yeo G, Black DL, Sun H et al. 2009. Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. *Mol Cell* **36**: 996-1006.
- Yabuki A, Tame A. 2015. Phylogeny and Reclassification of Hemistasia phaeocysticola (Scherff) Elbrachter & Schnepf, 1996. *J Eukaryot Microbiol* **62**: 426-429.
- Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**: 377-394.
- Yoon JH, Abdelmohsen K, Srikantan S, Yang XL, Martindale JL, De S, Huarte M, Zhan M, Becker KG, Gorospe M. 2013. LincRNA-p21 Suppresses Target mRNA Translation (vol 47, pg 648, 2012). *Mol Cell* **50**: 303-303.
- Yuan JH, Liu XN, Wang TT, Pan W, Tao QF, Zhou WP, Wang F, Sun SH. 2017. The MBNL3 splicing factor promotes hepatocellular carcinoma by increasing PXN expression through the alternative splicing of lncRNA-PXN-AS1. *Nat Cell Biol* **19**: 820-+.
- Záhonová K, Lukeš J, Dacks J. in review. Abundant marine diplomonads protists possess supposedly exotic membrane-trafficking machinery; Implications for LECA. *Proc Natl Acad Sci USA*
- Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe Nicholas M, Ule J. 2013. Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell* **152**: 453-466.
- Zhang X, Yang F, Zhan X, Bian T, Xing Z, Lu Y, Shi Y. 2023. Structural basis of pre-tRNA intron removal by human tRNA splicing endonuclease. *Mol Cell* **83**: 1328-1339 e1324.
- Zhang YJ, Qian JJ, Gu CY, Yang Y. 2021. Alternative splicing and cancer: a systematic review. *Signal Transduct Tar* **6**.
- Zhao Y, Liu Y, Zhang Q, Liu H, Xu J. 2023. The Mechanism Underlying the Regulation of Long Non-coding RNA MEG3 in Cerebral Ischemic Stroke. *Cell Mol Neurobiol* **43**: 69-78.
- Zheng CL, Fu XD, Gribskov M. 2005. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* **11**: 1777-1787.
- Zimmerly S, Semper C. 2015. Evolution of group II introns. *Mobile DNA-Uk* **6**.
- Zuckerman B, Ulitsky I. 2019. Predictive models of subcellular localization of long RNAs. *RNA* **25**: 557-572.