BALANCING SPLICING SPEED WITH FIDELITY:

ROLE OF A STRUCTURAL TOGGLE IN THE RNASEH DOMAIN OF PRP8

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Madhura Raghavan

August 2018

BALANCING SPLICING SPEED WITH FIDELITY:

ROLE OF A STRUCTURAL TOGGLE IN THE RNASEH DOMAIN OF PRP8

Madhura Raghavan, Ph.D.

Cornell University 2018

The molecular mechanisms by which the spliceosome achieves high fidelity during pre-mRNA splicing while simultaneously maintaining a high rate remain poorly understood. Here, I show a role for a core spliceosomal factor, Prp8, in balancing splicing speed with fidelity. An RNaseH-like domain (RH) within Prp8 contains a 17 amino acid insertion called the RH extension, which has been demonstrated to exist in two conformations: a β-hairpin, or an open loop structure. My work demonstrates that these two structures of the RH extension are associated with two distinct functional spliceosomal states. I demonstrate that mutations that increase the relative stability of the open loop conformation result in fast, but error-prone splicing. By contrast, mutations that increase the relative stability of the β-hairpin conformation result in hyper-accurate, but slow splicing. I propose a model where the RH extension toggles back and forth between the two conformations during splicing and helps balance speed with fidelity.

Further, to better understand the functional role of RH extension in splice site usage, I have investigated its evolution across organisms which display varying levels of degeneracy in their splice site sequences. Remarkably, the RH extension residues are invariant among

widely diverged species across almost all domains of eukaryotic life, the vast majority of which have degenerate splice site sequences. By contrast, in several organisms where the amino acid sequence of the RH extension has evolved, splice site sequences are seen to conform to a more rigid consensus, suggesting the fascinating possibility that the RH extension has driven the push towards constricting the splice site sequences present in these organisms.

BIOGRAPHICAL SKETCH

Madhura Raghavan was born in July 1989 in the coastal city of Chennai (formerly Madras), India. Growing up, the biology curriculum in her high school included primarily botany and zoology, and none of the cool molecular biology and genetics stuff that she does today. She developed a liking for botany and spent many evenings with friends collecting and examining plant specimens, something that Carolus Linnaeus would have found appealing in a potential graduate student in his lab.

When it came time for college, she wanted to pursue a course in biology and joined Anna University in her hometown for a Bachelors' degree in Industrial Biotechnology. It was here that she was introduced to the "new-age" late $20^{th}/21^{st}$ century biology. She found it fascinating and explored different areas of research in biology during her summer internships at the Indian Institute of Science, National Institute of Genetics, Japan and National Institute for Physiological Sciences, Japan. Graduating in 2010, she decided to pursue a Ph.D. and joined the Genetics, Genomics and Development Program at Cornell University in 2011. She began her doctoral work in Dr. Jeff Pleiss's laboratory and enjoyed working on mechanisms regulating mRNA splicing. Now she hopes to apply the skills she has learnt in her Ph.D. to valuable use and contribute to science in the years to come.

ACKNOWLEDGMENTS

I can’t thank my advisor, Jeff, enough for guiding me through my Ph.D. The reason I joined his lab was because of the comfort he created for a new graduate student to discuss her fledgling ideas without any inhibitions. He would guide me in the right direction without being dismissive and these interactions have helped me tremendously to grow as a scientist without being afraid of making mistakes. I have learnt much from his smart and creative approach to solving problems. I am grateful to him for providing me the space and time to design and troubleshoot experiments on my own. This has been instrumental in shaping my critical-thinking skills. The most impressive thing about Jeff is his optimistic attitude and this has helped me look past failures in experiments and keep moving on. Most of all, I am really thankful to him for always being available to discuss science and for creating a stimulating environment to do science.

I am much grateful to Dr. Andrew Grimson and Dr. Eric Alani for their interest in my growth as a scientist. I could not have found more supportive committee members and they have been instrumental in keeping me on track. I have always appreciated their taking the time to come and enquire after my progress and future career plans and give valuable inputs to my research. I would like to thank Dr. John Lis for his valuable scientific insights. I have been inspired by his ideas looking at finer mechanistic details to connect the dots for the larger mechanism.

I am grateful to our collaborators, Dr. Christine Guthrie, Dr. John Abelson and their lab members at UCSF, for the work on Prp8. This collaboration was to me an opportunity to learn new ways of thinking about biology, delving into intricate mechanistic details. It has immensely helped me in learning to find the most important and interesting questions to address. I would also like to thank Dr. Megan Mayerle at UCSF for the many insightful interactions on the project. I am also grateful to Dr. Thomas Cleland and Dr. Michelle Tong at Cornell for the opportunity to collaborate with them on a project to dissect gene expression networks involved in the learning and memory in the mouse brain.

My time in the lab would not have been so enjoyable without the awesome current and past members of the Pleiss lab. They have been outstanding people and I can't thank them enough for making the lab atmosphere very positive, collaborative and interactive. I have learnt much from them, importantly, to take active interest in each other's projects and help and share knowledge with each other without hesitation. I hope I was able to reciprocate. I would like to thank Hansen with whom I had fun working on the Mud1 project and Laura, whom I looked up to as a mentor during my initial years in the lab. I would like to thank Ben for the many enjoyable chatting sessions on science and for critically evaluating my ideas for experiments, Nick for help with lariat sequencing, Zach for being the unofficial IT person in the lab, ever-ready to help, and helping set up servers for everyone to use and Mike for the critical feedback on my project. I would also like to thank the Grimson and Kwak labs for giving me important inputs on my project.

I owe everything to the love and support of my parents, Vidhya and Raghavan, and sister, Mekhala. They have been instrumental in instilling in me the love for learning. I hope I have serendipitously conveyed my thanks to my father, Raghavan, in the form of giving publications that will be cited in his name. I am especially very thankful to my sister who was my mentor during my undergrad days and even during grad school. I am grateful to my grandparents for constantly reminding me to pursue a career keeping societal interests in mind and not just my own. Life wouldn't have been so enjoyable and meaningful in Ithaca without all the amazing friends that I have made here as well as those from before with whom I have stayed in touch. Those in Ithaca have especially made this a home away from home.

In the end, I would like to acknowledge the privilege that helped me grow up in an environment that provided the exposure to all these opportunities and helped me get equipped with the skills necessary to pursue something I was interested in.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The pre-mRNAs of most eukaryotic genes must undergo a process termed splicing, wherein intervening, noncoding introns are removed from the pre-mRNA to generate a properly translatable mRNA (Sharp, 2005). Splicing must occur with high precision, as activation of incorrect splice sites will generate aberrant mRNAs which might well be detrimental to the organism. Indeed, recent studies have implicated around one third of the single nucleotide polymorphisms (SNPs) associated with human genetic diseases as affecting the splicing pathway (Singh & Cooper, 2012). However, it is generally thought that achieving high accuracy in biological systems comes at a cost of reduced speed. So, the challenge for the spliceosome is to maintain accuracy while still processing the pre-mRNAs efficiently to meet cell's needs. Even though the locations of introns have been mapped in many organisms and much is known about the basic molecular mechanisms involved in splicing, the molecular basis by which high fidelity is maintained without compromising speed remains unclear. Therefore, the fundamental question that I have addressed in my thesis is this: how does splicing occur with both high efficiency and accuracy?

SPLICING BASICS

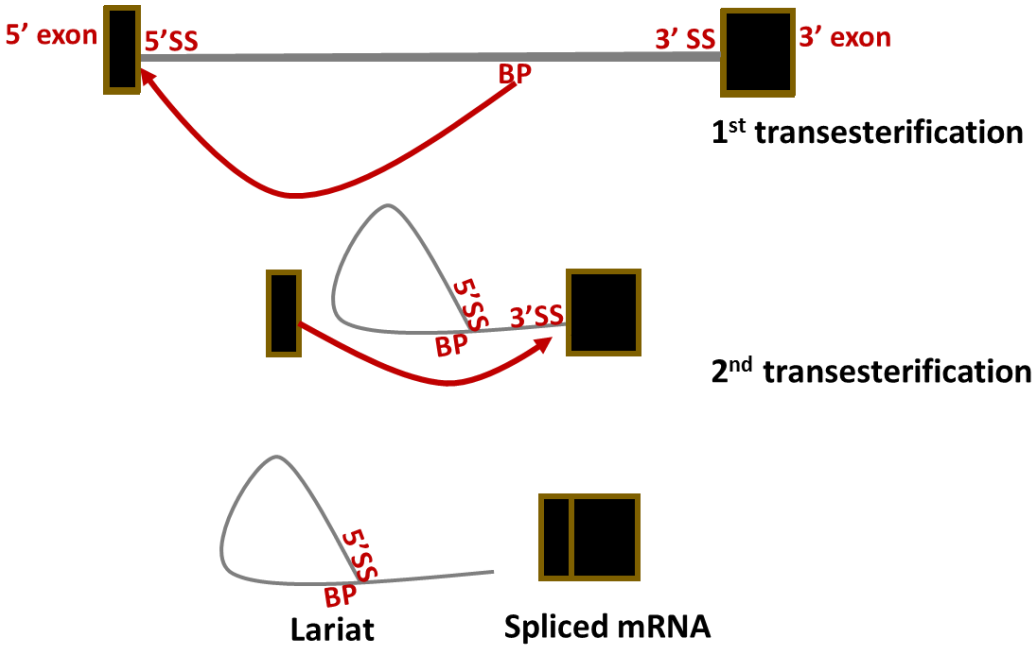
It is widely accepted in the field that the spliceosome, the macromolecular machine that catalyzes the splicing reaction, is not a pre-assembled complex but rather assembles anew on each substrate, with components added and removed in a regulated, step-wise assembly (for reviews, please see (Fica and Nagai 2017; Scheres and Nagai 2017; Wahl, Will, and Lührmann 2009; Will and Lührmann 2011)). After initial intron recognition and spliceosome assembly, the catalytic site is generated, allowing splicing to occur in two transesterification reactions (Figure 1.1). While the two reactions themselves are straight forward chemical

reactions aided by Mg^{2+} ions in the spliceosome's catalytic site (Fica et al., 2013; Steitz & Steitz, 1993), it is the act of identifying the appropriate substrates and positioning them correctly for catalysis that is a challenge for the spliceosome.

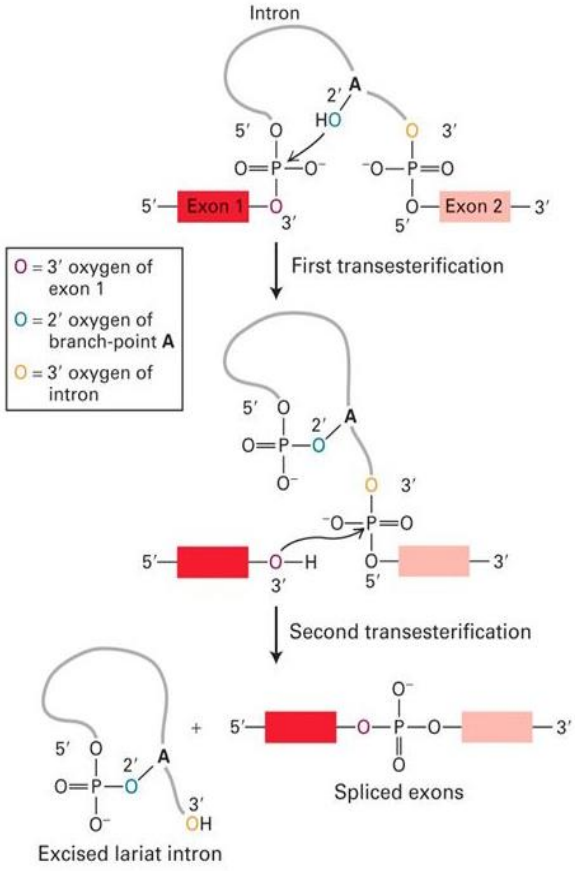
Figure 1.1 Splicing removes introns from the pre-mRNA through two transesterification reactions.

a. There are three main elements in the intron (grey line) that participate in splicing: the 5'splice site (5'SS), the branch point (BP) and the 3'splice site (3'SS). In the first reaction, the 2'-OH of the adenosine at the branch point reacts via a transesterification reaction with the phosphodiester at the 5'splice site, resulting in a free 3'-OH on exon 1 and a lariat linkage within the intron. In the second reaction, the free 3'-OH at the end of exon 1 reacts via a transesterification with the phosphodiester at the 3' splice site resulting in the ligation of the two exons (boxes). b. The transesterification reactions at the two steps are shown (figure reproduced from *Pre-mRNA splicing, Lecture 1, Douglas Black, http://slideplayer.com/slide/6905003/*).

a.


5' exon
5'SS
3' SS
3' exon
BP
1st transesterification
5'SS
3'SS
BP
2nd transesterification
5'SS
BP
Lariat
Spliced mRNA

b.


Intron
2'
A
HO
5'
O
3'
Exon 1
Exon 2
First transesterification
O = 3' oxygen of exon 1
O = 2' oxygen of branch-point A
O = 3' oxygen of intron
Second transesterification
Spliced exons
Excised lariat intron
OH

The spliceosome is composed of five small nuclear ribonucleoprotein complexes (snRNPs) and scores of auxiliary proteins. Initial recognition of the two substrates of the chemical reaction, namely the 5'SS and the BP, is carried out largely through basepairing with the U1 snRNP and U2 snRNP, respectively, with additional help from the branchpoint binding complex (BBC). While binding of these components are important for early assembly steps, most of these components are no longer present in the spliceosome when catalysis occurs (Scheres and Nagai 2017). By contrast, the U4/U5/U6 tri-snRNP and the Prp Nineteen Complex and Nineteen Related complex (NTC/NTR) only stably join after assembly of the U1 and U2 snRNP complexes. Structural rearrangements take place upon tri-snRNP addition that result in the formation of the active site and positioning of the substrates for catalysis (Figure 1.2). While the activity of the early acting U1 and U2 snRNPs can clearly have important impacts on the fidelity of the splicing pathway, it has also been shown that the spliceosome can proofread substrates at later steps in the pathway (Koodathingal & Staley, 2013; Query & Konarska, 2006). Therefore, to understand how splicing balances speed and fidelity, it is clear that both initial assembly and splicing catalysis must be considered.

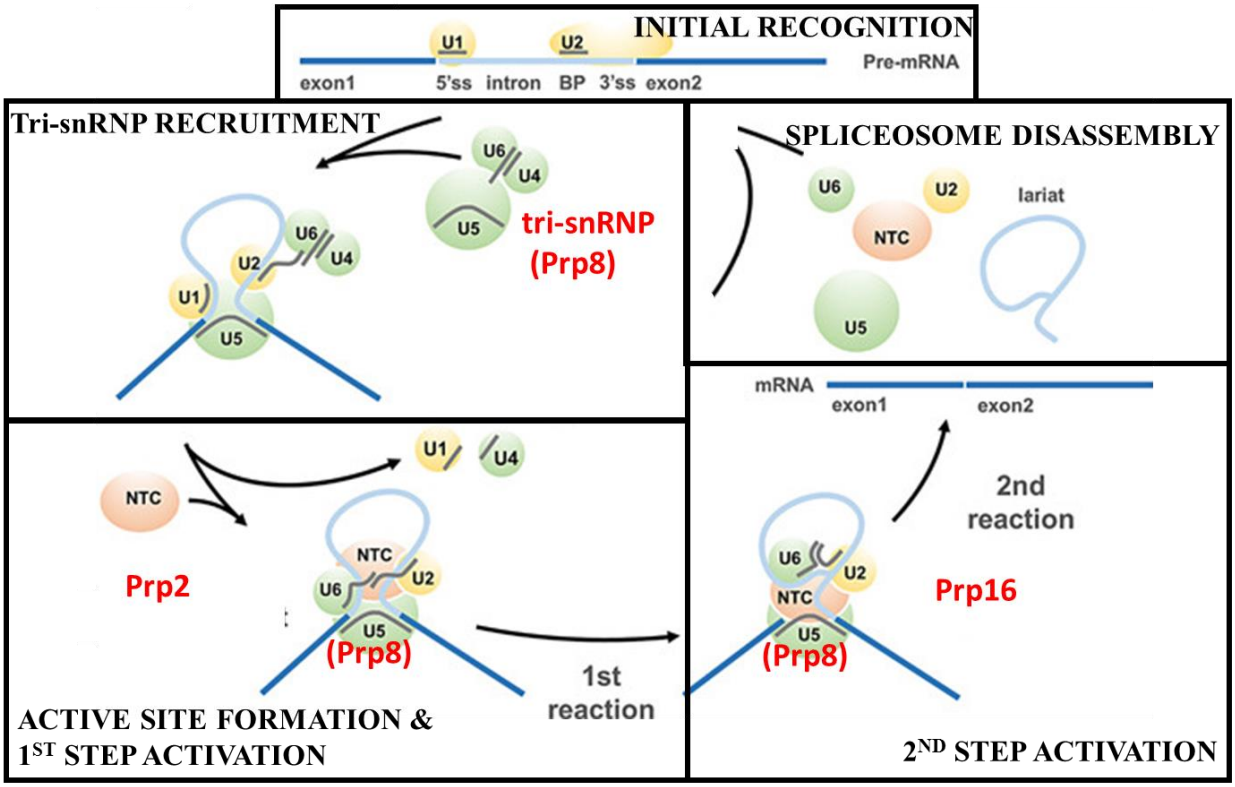


Figure 1.2: Splicing cycle (Modified from (Martinho, Guilgur, & Prudêncio, 2015)**)**

The 5'SS and BP-3'SS sequences are initially bound by U1 and U2 snRNPs, respectively. Following recruitment of the U4/U5/U6 tri-snRNP, several regulated changes in structure and composition are undertaken. The U1 and U4 snRNPs are released from the complex, allowing for formation of the active site, which is positioned by the protein Prp8, a stable component of the U5 snRNP. The activity of the RNA helicase Prp2 is required in order for 1st step chemistry to occur. The spliceosome is then activated for the 2nd step by another RNA helicase, Prp16. Following exon ligation, the spliceosome is disassembled.

Active site formation

After initial assembly of the U1 and U2 snRNPs onto the pre-mRNA, the U4/U5/U6 tri-snRNP is recruited to the spliceosome, components of which will eventually form the catalytic site. Structural and compositional rearrangements are then catalyzed, with factors being recruited and displaced at distinct steps, aided by ATP-driven helicases (Scheres and Nagai 2017). Presumably, two main tasks drive the spliceosomal rearrangements at this juncture: (1) formation of the active site that catalyzes the two transesterification reactions, and (2) ensuring that the right substrates are chosen for catalysis and exposed at the right time.

Recent high resolution structures of the spliceosome show that the active site forms before the first step and is identical at the two steps of splicing (Fica and Nagai 2017; Scheres and Nagai 2017). The active site rests on a cavity formed by the largest spliceosomal protein Prp8. The active site, as predicted by previous cross linking and biochemical studies, is made up of the intramolecular stem loop (ISL) of the U6 snRNA, helix I of the U2/U6 duplex, loop I of the U5 snRNA, and magnesium (Mg^{2+}) metal ions (Figure 1.3). One of the first step substrates, the 5'SS, is partially basepaired to the U6 snRNA and is positioned in the active site, while the BP region, whose bulged adenosine is the other first step nucleophile, is bound by the U2 snRNA (termed as branch helix) and is not brought into the active site until just before the 1st step. The 3' end of the first exon is bound by stem 1 of the U5 snRNA, which along with the 5' stem of the U6 snRNA and the 5'-exon are bound to the N-terminal domain of Prp8.

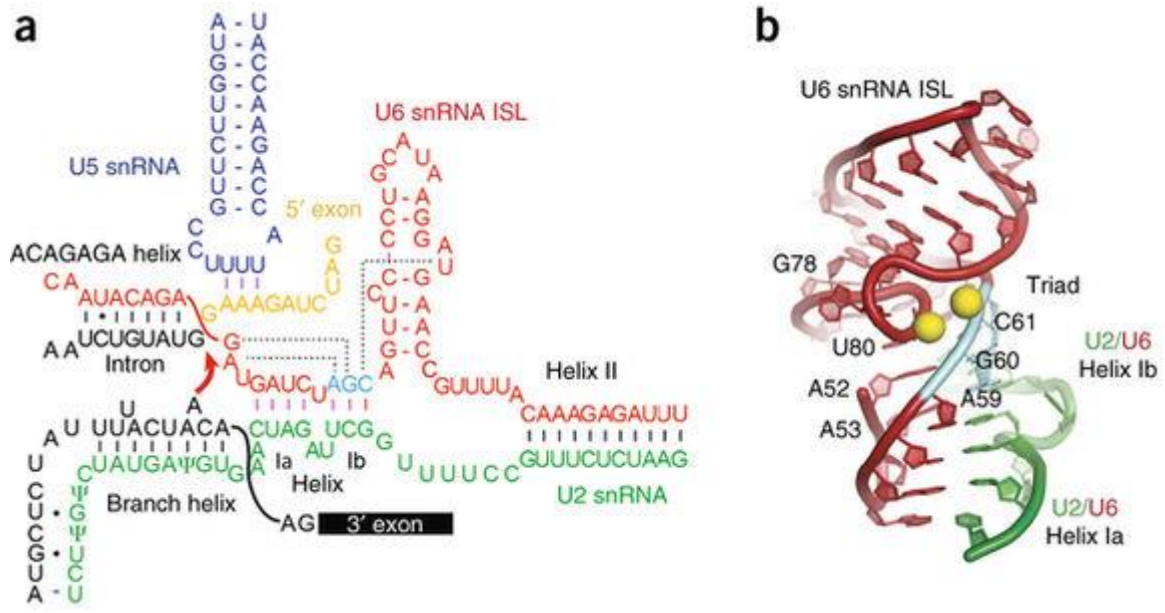


Figure 1.3 – Catalytic RNA core of the spliceosome (reproduced from (Fica and Nagai, 2017)**)**

a) The catalytic RNA core prior to the 1st chemical step. The active site is formed by the U2/U6 helix I (Ia and Ib), U6 snRNA ISL, and loop I of U5 snRNA. The 5'SS base pairs with U6 snRNA, while the 5' exon is held by U5 snRNA. The BP region basepairs with U2 snRNA to form the branch helix. Here, the BP-adenosine does not base pair but rather is bulged out, exposing the 2'OH. The red arrow indicates the 1st transesterification reaction from the BP-adenosine to the guanosine at the first position in the 5'SS. (**b**) The three-dimensional structure of the active site RNA following the 1st chemical step is shown. A triplex forms between U2/U6 helix I and U6 snRNA ISL and three key residues are involved in coordinating the Magnesium ions necessary for catalysis.

Positioning of substrates for catalysis

While the active site remains unchanged at the two steps, the substrates are moved around for the two steps of splicing (Figure 1.4). The nomenclatures for the different spliceosomal complexes are described below, while the different rearrangements that occur are described in detail.

B complex – 5'splice site is paired with U6 snRNA and positioned in what will become the active site

B^{act} complex – Spliceosome where the active site made of U2/U5/U6 snRNAs has formed.

B* complex – Spliceosome activated for 1^{st} step catalysis.

C complex – Just after 1^{st} chemical step (lariat intermediate formation)

C* complex – Spliceosome activated for 2^{nd} step catalysis

ILS – Just after 2^{nd} chemical step (exon ligation)

For B complex formation, the pairing of the U1 snRNA with the 5'SS is disrupted by the helicase Prp28, allowing for transfer of the 5'SS to the U6 snRNA. The U6 snRNA base paring with the 5'SS will eventually become a part of the active site (Plaschka, Lin, and Nagai 2017; Wahl, Will, and Lührmann 2009)

In the B^{act} complex, although the active site has formed, the substrates are away from each other and cannot participate in the first chemical step. Whereas the 5'SS is already in the active site and held by the U6 snRNA, the branch helix is held by the SF3 complex with the BP adenosine being bound by Hsh155 and located approximately 50 Å away from the catalytic center (Rauhut et al., 2016; C. Yan, Wan, Bai, Huang, & Shi, 2016).

The B* complex structure is not yet available. Instead, the C complex, which forms right after the 1^{st} transesterification, can be used to infer the structure of the B* complex.

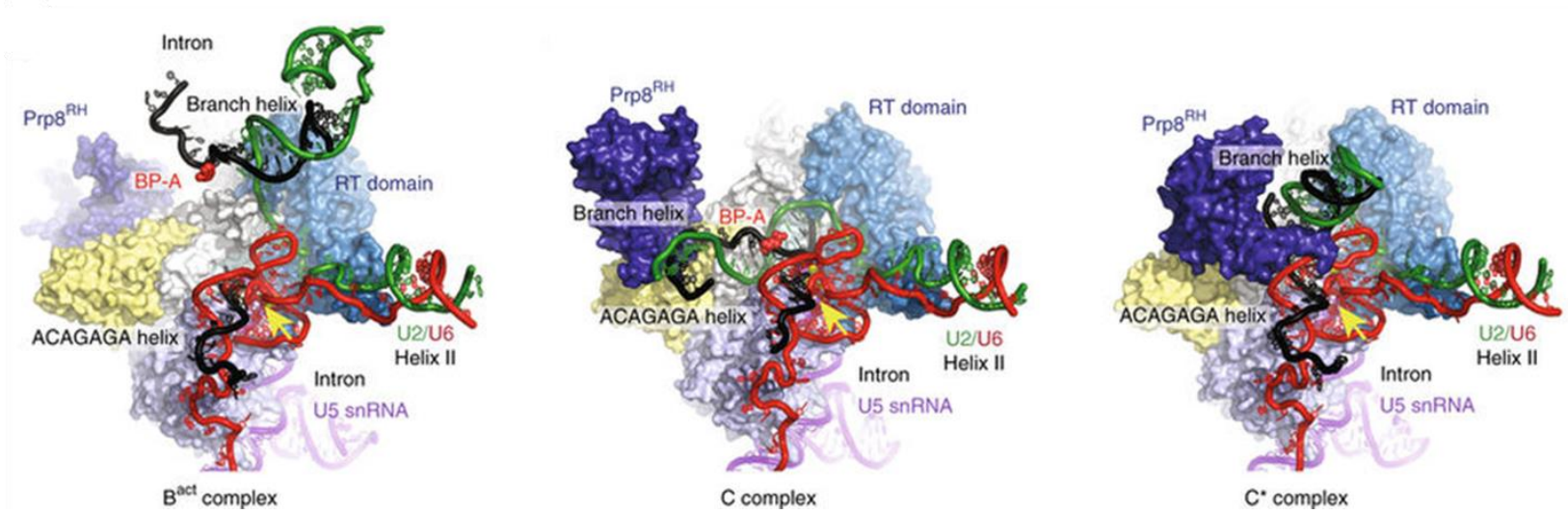


Figure 1.4: Positioning of substrates for the two chemical steps (reproduced from Fica & Nagai, 2017)

The catalytic RNAs and intron are color coded and the labels indicate their color. The catalytic center rests on Prp8. The different domains of Prp8 are colored as follows: the N-terminal domain in light blue, the RT domain in aquamarine, the linker domain in white, the EN domain in light yellow, and the RNaseH (RH) domain in deep blue. The active site is unchanged at the two steps, while the substrates are moved in and out of it. The branch helix, which holds the 1st step nucleophile BP-adenosine (BP-A), is away from the active site in $B^{act.}$ However, in the C complex, which can be taken to reflect the state at B*, the branch helix has moved to position the BP-adenosine in the active site. The 5'SS is already positioned in the active site by the ACAGAGA region of U6 snRNA in these two complexes. In the C* spliceosome, which is activated for 2nd step catalysis, the branch helix has moved again to allow the 2nd step substrate, 3'-exon, into the active site.

In the C complex (Galej et al., 2016; Wan, Yan, Bai, Huang, & Shi, 2016), the SF3 complex is dislodged from the branch helix and Hsh155 no longer binds the BP adenosine. The branch helix has undergone a rotation so that the nucleophilic 2'OH of the bulged adenosine sits in the active site, having taken part in the nucleophilic attack. The substrates are held in position by first step factors.

In the C* complex, rotation of the branch helix moves the BP adenosine 20 Å away from the catalytic center allowing room for the incoming 3′-exon for exon ligation. The substrates are held in position by second step factors (Fica et al., 2017; Chuangye Yan, Wan, Bai, Huang, & Shi, 2017).

The above transitions that involve positioning of the substrates in and out of the active site are mediated by highly coordinated structural changes in the components around the active site and induced by RNA helicases (De, Schmitzová, & Pena, 2016) that bind to the pre-mRNA downstream of the branch helix away from the active site (Scheres and Nagai 2017). It is thought that the structural rearrangements driven by the helicases give directionality to the splicing process as they are ATP-consuming reactions for which no evidence exists for reversibility (for reviews, see (Cordin & Beggs, 2013; De et al., 2016; Liu & Cheng, 2015)). It is unclear how information from the active site and factors bound to the substrates is transmitted to the helicases to activate them at the right time for catalysis, as premature activation before the substrates and catalytic core are proofread might lead to incorrect splicing. It is conceivable that the structural geometry of the active site with the correct substrates induces changes in the surrounding factors that then communicate with the helicases to accelerate forward activation steps in the pathway. Alternatively, a non-optimal structural geometry resulting from usage of incorrect substrates might delay the subsequent

steps, eventually leading to discard of those sites.

PRP8

Of the factors surrounding the active site with a potential role in transducing information from the active site as well as influencing active site geometry, one of the most important is Prp8. A stable component of the U5 snRNP, Prp8 is the most highly conserved protein in the spliceosome (Grainger & Beggs, 2005). This conservation suggests a critical role for Prp8 in splicing and indeed, functional genetic and biochemical studies in the last couple of decades had placed Prp8 near the catalytic core of the spliceosome and the recent high resolution structures confirm this. The structures make it clear that the spliceosome is a protein-directed ribozyme, where protein components are essential to direct the formation and activation of the catalytic site, with Prp8 playing arguably the most important of these roles (Shi, 2017), making a scaffold that holds the active site of the spliceosome (Figure 1.4) (Fica and Nagai 2017; Shi 2017). In fact, until high-resolution structures were able to show that the snRNAs formed the catalytic center, it was not clear if Prp8 or the snRNA network was the enzyme that catalyzed splicing (Abelson, 2008; Guthrie & Collins, 2000). Thus, Prp8 plays a central role in the formation and activation of the active site and owing to its position has potential for influencing both the fidelity and efficiency of catalysis.

Prp8 genetically interacts with activation factors as well as the catalytic RNAs, and can suppress defects in splice site mutations

Not surprisingly, given its critical location, genetic studies have implicated a role for Prp8 in key steps of spliceosomal activation (for review, see (Grainger & Beggs, 2005)). Alleles

of *prp8* that suppress/exacerbate defects in steps necessary for the formation of the active site have been described. For instance, defects in U4/U6 unwinding leading to catalytic site formation in the *U4-cs1* mutant are suppressed by mutations in Prp8 (Kuhn & Brow, 2000). Also, Prp8 has been shown to genetically interact with regions at and around the active site: region of U6 snRNA involved in the formation of the ISL (Kuhn, Reichl, & Brow, 2002), region of U2 snRNA involved in forming U2/U6 helix I and ISL I (Xu et al., 1998) as well as loop I of U5 snRNA (Frank, Patterson, & Guthrie, 1992). Further, it also genetically interacts with the helicases Prp16 (Hotz and Schwer 1998; Query and Konarska 2004; Umen and Guthrie 1995) and Prp22 (Schneider, Campodonico, & Schwer, 2004) which drive ATP-fueled remodeling events before and after the 2nd transesterification reaction. Prp8 has also been shown to suppress mutations in either the 5'SS, 3'SS or BP sequences which lead to defects in either the first or second steps of splicing (Collins and Guthrie 1999; L. Liu, Query, and Konarska 2007; Query and Konarska 2004; Siatecka, Reyes, and Konarska 1999; Umen and Guthrie 1996). Suboptimal substrates that normally splice poorly are used well in spliceosomes containing mutations in Prp8. Therefore, given that it is a core component shown to regulate both catalysis and fidelity, a central role for Prp8 in maintaining a balance between the two would not be surprising.

HOW MIGHT PRP8 BE LINKED TO BALANCING SPLICING SPEED AND FIDELITY?

Does the RH extension in Prp8's RNaseH domain link splicing speed with fidelity?

Prp8 has four domains: the N-terminal, Large (consisting of RT, Endonuclease and linker domains), RNaseH (used interchangeably with RNaseH-like as it does not possess the active

site residues essential for RNaseH activity (Pena, Rozov, Fabrizio, Lührmann, & Wahl, 2008; Ritchie et al., 2008; Yang, Zhang, Xu, Heroux, & Zhao, 2008)) and Jab1/MPN domains, connected by flexible linker peptides (Galej, Oubridge, Newman, & Nagai, 2013). The active site of the spliceosome is held by the N-terminal and Large domains(Fica and Nagai 2017), while the RNaseH and Jab1/MPN domains have been implicated as regulators of some key structural rearrangements of the spliceosome towards catalytic activation (Absmeier, Santos, & Wahl, 2016) (Figure 1.4). Although all domains are crucial for splicing, functional studies show the importance of the RNaseH domain in ensuring correct splicing. Many of the *prp8* alleles mentioned above that either suppress splice site mutations or genetically interact with key activation factors/RNA elements map to the RNaseH domain (Figure 1.5), with its RH extension being a hot spot (Grainger & Beggs, 2005). The RH extension is a 17 amino acid unusual protrusion that projects out of the domain. Earlier studies had predicted that such a protrusion could likely interact with RNA elements, which is now proved true by the recent structures (Pena et al., 2008; Ritchie et al., 2008; Yang et al., 2008). The importance of the RH extension is further underscored by the observation that deletion of the extension is lethal in yeast (Mayerle et al., 2017).

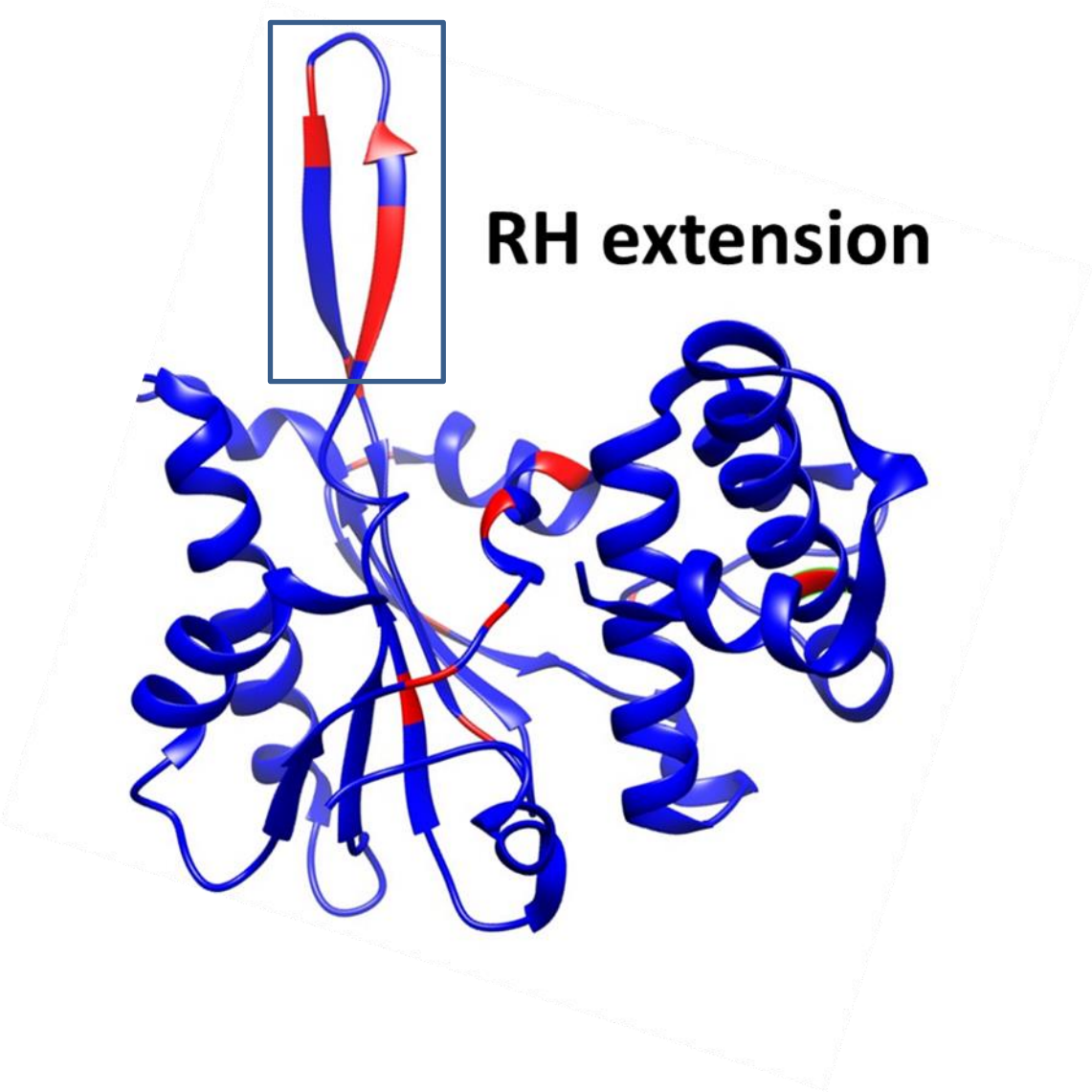


Figure 1.5: Locations of suppressor mutations in the RNaseH domain of Prp8

The locations of suppressor mutations are shown in red in the RNaseH domain of Prp8. RH extension is the finger protruding from the domain and is highlighted by the box. The structure is adapted from *PDB ID: 3E9O* (Pena et al., 2008) and the mutations are as referenced in (Grainger & Beggs, 2005).

Recent structural studies reveal interesting dynamics of the RNaseH domain and its extension during splicing. While the active site itself is unchanged and sits fixed on the N-terminal and Large domains of Prp8 during both steps of splicing, the RNaseH domain moves about the active site (Fica & Nagai, 2017), potentially facilitating structural transitions necessary for the splicing cycle (Figure 1.6).

The formation of B^{act} spliceosome involves the unwinding of U4/U6 snRNA duplex by the helicase Brr2 to release the U6 snRNA for interacting with U2 snRNA to form the catalytic RNA core. Before B^{act} formation, the RNaseH domain binds to U4/U6 snRNA and blocks Brr2's interaction with U4 snRNA, thereby negatively regulating Brr2 activity (Mozaffari-Jovin et al., 2012). This presumably precludes premature formation of the catalytic RNA core before proofreading of the 5'SS/U6 snRNA pairing. The formation of the B* complex after B^{act} involves displacement of SF3 factors, including Hsh155, from the branch helix by the helicase Prp2.

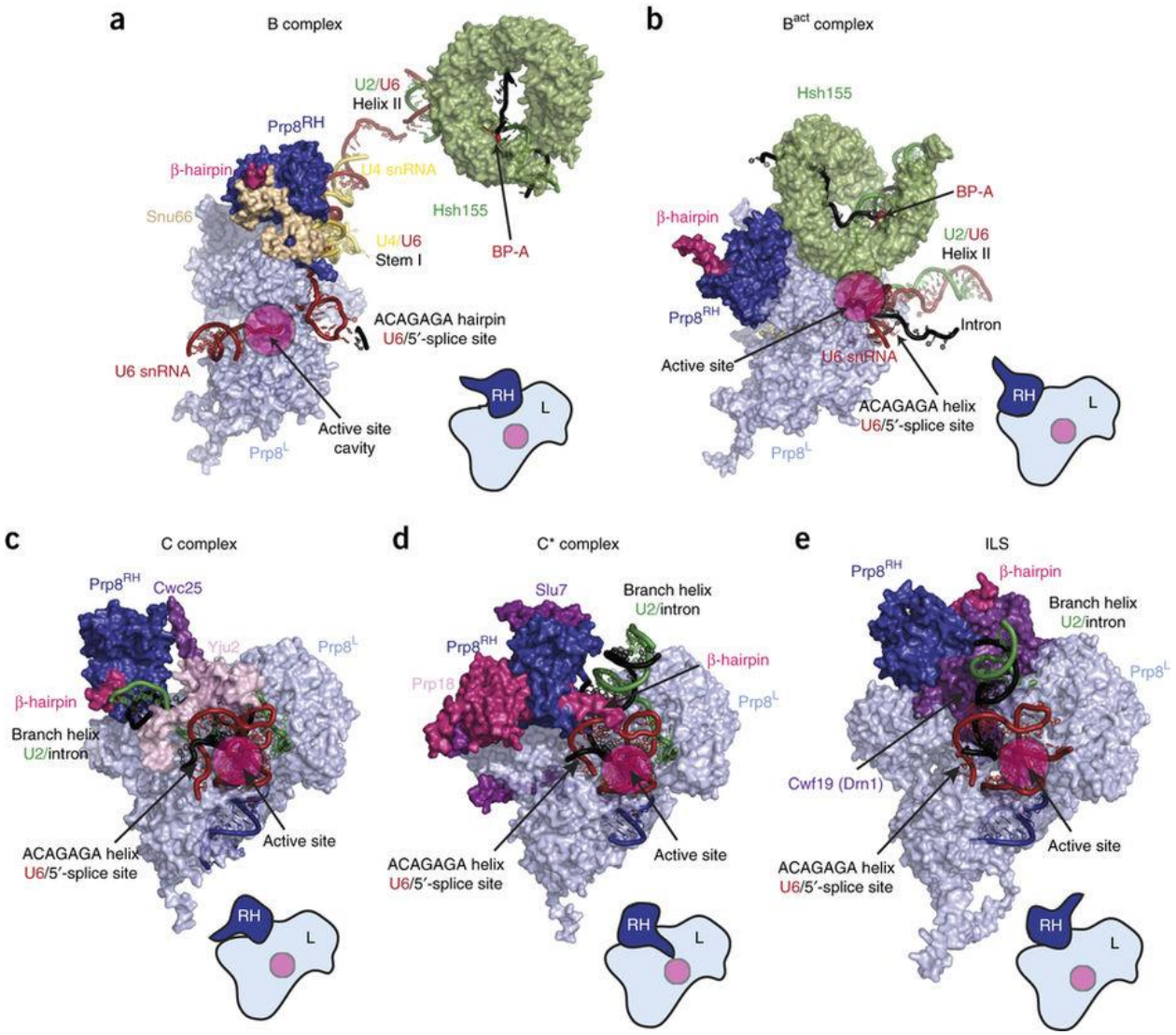


Figure 1.6 Movement of Prp8 RNaseH domain during splicing (reproduced from Fica and Nagai, 2017)

Prp8's RNaseH domain (in deep blue) moves with respect to the Large domain of Prp8 (grey), which holds the active site (pink circle). At each step, it interacts with factors that are involved in remodeling at that step. β-hairpin refers to the RH extension in the RNaseH domain of Prp8 and is colored in magenta.

In the B^{act} complex, the RNaseH domain moves from its earlier position near the U4/U6 duplex and instead resides close to Prp2 and Hsh155, although apparently without directly interacting with either (Yan et al., 2016). Given its proximity, it is conceivable that positioning of the RNaseH domain could impact Prp2 and its activation.

In the C complex, which is formed after the first chemical step, it is seen that the branch point adenosine which was released by Prp2 action, is in the active site and is held in position by first step factors Cwc25 and Yju2: two factors with which the RNaseH domain interacts. Furthermore, the RH extension holds the branch helix which extends out of the active site (Galej et al., 2016; Wan et al., 2016).

In C* complex, when the spliceosome has been activated for the 2^{nd} chemical step by the helicase Prp16, the branch helix has been pulled from the active site to allow room for the 3'-exon for exon ligation and this conformation is stabilized by the second step factors Prp18 and Slu7. The RNaseH domain has moved again and is interacting with Prp18 and Slu7. Also, the RH extension in RNaseH domain now sits in the groove between the U6/5'SS helix and branch helix, presumably stabilizing that structure after they have been pulled out of the active site (Fica et al., 2017; Yan et al., 2017).

Thus, structurally, the RNaseH domain and the RH extension occupy places at the heart of structural rearrangements at each step and potentially contribute to and/or regulate these rearrangements.

Interestingly, the suppressor mutations that map to the RNaseH domain and RH extension have two opposing genetic phenotypes, which were first proposed in 2004 by Query and Konarska to stabilize the first and second step catalytic conformations of the spliceosome (Query & Konarska, 2004). Later, crystallographic studies of the RNaseH

domain in 2008 by the Zhao, MacMillan, and Luhrmann groups showed two structures in the asymmetric unit. The two structures were highly similar except for the RH extension, for which two conformations were observed: a rigid and structured β-hairpin structure or a more flexible open loop structure (Figure 1.7) (Pena et al., 2008; Ritchie et al., 2008; Yang et al., 2008). In 2013, the MacMillan group showed for a subset of *prp8* alleles that those that preferentially stabilize either the β-hairpin or open loop forms of the RH extension correlated with distinct genetic phenotypes that were previously proposed to be associated with1st step or 2nd step catalytic conformations of the spliceosome (Schellenberg et al., 2013b). All these remarkable studies pointed to two opposing phenotypes of the two conformations of the RH extension. At the time the Query and Konarska model was proposed, it was unknown how the catalytic site of the spliceosome looked at the two steps of splicing. However, the recent spate of high resolution cryo-EM structures in the last couple of years show that the catalytic conformation stays the same at both steps of splicing and this is inconsistent with the Query and Konarska model (Fica & Nagai, 2017).

Later in the chapter, I will propose that the two conformations of the RH extension are associated with catalytically silent and active conformations of the spliceosome that link splicing speed with fidelity. This model is based on our understanding of the ribosome and how it uses conformational toggling to balance speed and fidelity (Ogle & Ramakrishnan, 2005).

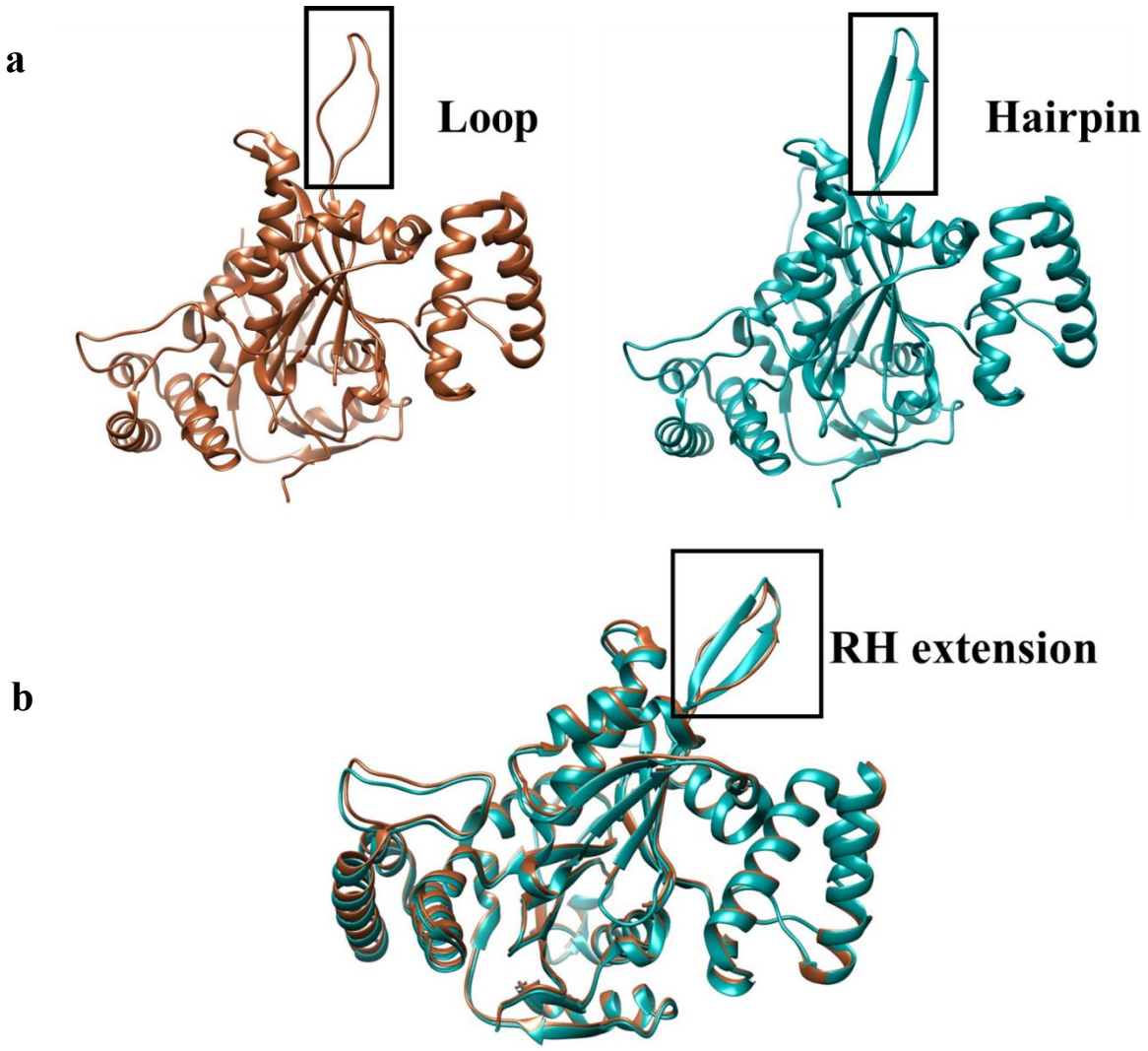


Figure 1.7 Loop and Hairpin structures of the RH extension

The two structures obtained in the crystal structure of Prp8's RNaseH domain shown separately (a) and superimposed on each other (b). RH extension (in black box) shows two conformations while other regions are highly similar in the two structures. Adapted from PDB ID: 4jk7 (Schellenberg et al., 2013).

RIBOSOME - CONFORMATIONAL TOGGLING LINKED TO BALANCING OF SPEED AND FIDELITY

The concept of structural changes being associated with the fidelity and speed of a biological process is not completely new, but rather has been well characterized for the ribosome (Figure 1.8), which has been described as toggling between open and closed conformations (Ogle, Murphy, Tarry, & Ramakrishnan, 2002; Rodnina, Fischer, Maracci, & Stark, 2017). In the open conformation, the A site is vacant and accepts a tRNA for addition of the next amino acid in the peptide. If a cognate tRNA is introduced, extensive basepairing exists between the codon and anti-codon, and the rate of dissociation of the tRNA is quite low. Further, the cognate codon/anti-codon basepairing induces interactions with the A site that accelerate structural changes in the ribosome that lead to the closed conformation. The closed conformation then accelerates irreversible GTP hydrolysis by EF-Tu, leading to the next steps in peptide bond formation. For cognate tRNAs, the forward process is accelerated through an irreversible GTP hydrolysis step, thereby ensuring high efficiency with optimal substrates. Alternatively, near cognate tRNAs by definition present a mismatched base at the A site in the open conformation, increasing the rate of dissociation. Moreover, the local structure generated by near- (or non-) cognate codon/anti-codon structures fail to stimulate GTP hydrolysis slowing the conformational rearrangement to the closed state but increasing the likelihood the tRNA is discarded. As such, a near-/non-cognate tRNA is dissociated before being taken to the next step in the pathway through a slower forward step, thereby ensuring high fidelity of the process. Thus, through conformational toggling, the ribosome balances both speed and fidelity. Stabilizing the closed conformation artificially by mutations (called the ‘Ram’ mutations) or with chemicals makes translation error-prone. In contrast,

stabilizing the open conformation by mutations (called the ‘Restrictive’ mutations) makes translation hyper-accurate (Ogle & Ramakrishnan, 2005).

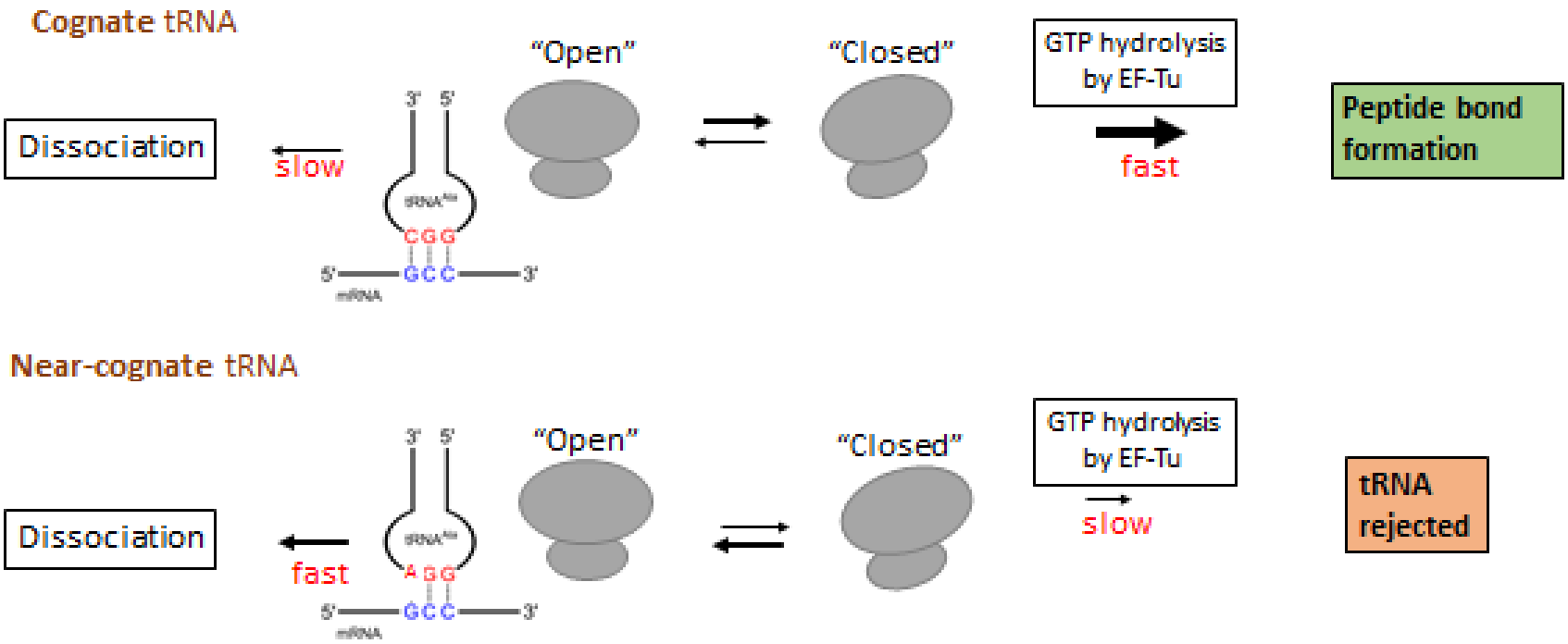


Figure 1.8: Conformational toggling linking speed and fidelity in the ribosome

Does such conformational toggling between fast/active and slow/silent states exist in the spliceosome and if so, is it associated with any irreversible energy-driven forward steps in the spliceosome? As described earlier, helicases perform ATP-mediated remodeling steps that drive splicing forward. Moreover, as described below, the existence of catalytically active and silent states in the spliceosome is predicted by structural studies of Group II introns, the ancestral version of the modern spliceosome.

GROUP II INTRON PREDICTS TWO STATES, CATALYTICALLY ACTIVE AND SILENT STATES, FOR THE SPLICEOSOME

Group II introns are self-splicing ribozymes found in many organisms in all domains of life and are widely believed to be the evolutionary ancestor of the spliceosome (Lambowitz & Zimmerly, 2011). They perform splicing through two SN_2 transesterification reactions similar to the mechanism in spliceosomal introns involving branching at the 1st step, where a bulged adenosine near the 3' end of the intron, or water, acts as a nucleophile and makes an attack at the 5'SS forming a 2'-5' phosphodiester linkage, and excision of the intron as a lariat at the 2nd step (Pyle, 2016). Earlier biochemical and functional genetic studies indicated many similarities between the spliceosome and the group II intron: the bulged adenosine is present in both, interactions between the exons and spliceosomal U5 snRNA (loop 1) are similar to the exon binding sequence–intron binding sequence interactions in group II introns and the intramolecular stem-loop (ISL) of U6 snRNA is highly similar to domain V (DV) in group II introns (Galej, Toor, Newman, & Nagai, 2018).

In 2008, the first crystal structure of a group II intron was obtained and this was followed by functional studies in spliceosomal introns to supplement previous knowledge on the

residues involved in metal ion binding based on group II structure (Toor, Keating, Taylor, & Pyle, 2008). In 2012, structures of the group II intron at different stages were captured and these suggested that the group II intron uses the same active site during both steps (Marcia & Pyle, 2012). At that time, high resolution structures of the spliceosome were still not available and it was unclear whether the spliceosome possessed the same or different active sites for the two steps. The widespread model in the field at the time was the presence of two different active sites at the two steps and this was widely applied in the understanding of the *prp8* alleles as well following the model described by Query and Konarska. However, given the high similarity in the mechanism and RNA interaction network of group II introns and the spliceosome, I hypothesized that the active site of the spliceosome is also the same at the two steps. This is now confirmed by recent high resolution structures of the spliceosome. Further, structural data suggested that the group II intron passes through an intermediate transitional structure between the two catalytic steps (Figure 1.9). In this state, the catalytic metal ion center (both the spliceosome and group II introns require metal ions for splicing) is disrupted, suggesting that this is a "silent" intermediate state (Marcia & Pyle, 2012). Although a "silent" state has not been observed for the spliceosome, considering that it is thought to have evolved from group II introns, it is very likely that there are two such states for the spliceosome as well and I hypothesize that two such states exist for the spliceosome.

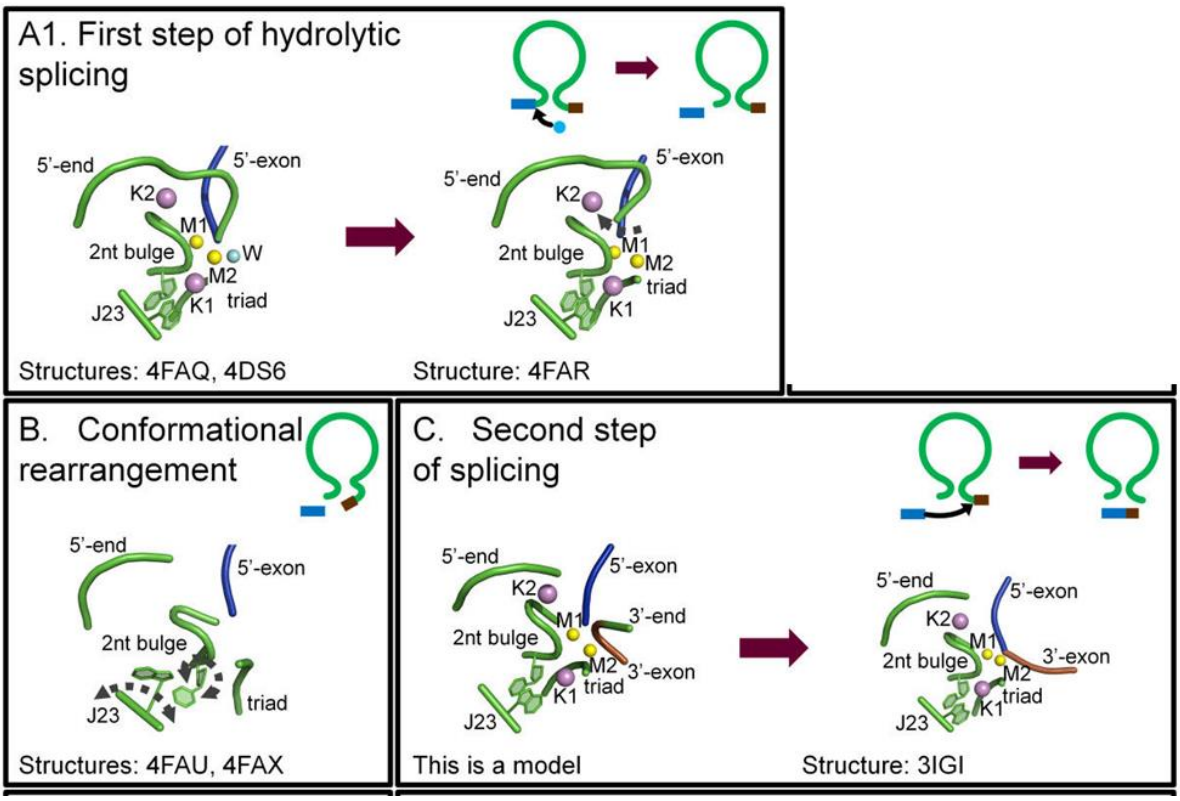


Figure 1.9: Group II intron shows a catalytically silent state in between the two steps (reproduced from *Marcia et. al, Mobile DNA, 2013*)

The structures at the first (a), second (c) steps of group II intron splicing as well at the intermediate step between the two steps (b) are shown. The first step of splicing involves a nucleophilic attack by a water molecule on the 5'SS, resulting in the release of the 5'-exon. The second step is similar to the reaction by the spliceosome. The structures show the active site and the reactants. 5'-exon is in blue, water molecule in cyan, 3'-exon in black, intron in green, metal ions in yellow and purple. The active site conformation (triad, J23, 2nt bulge) is similar in the first and second steps of splicing. However, there is a slight conformational change in the intermediate step, where the active site does not bind to metal ions. Since metal ions are essential for catalysis, this presumably reflects a catalytically silent state.

TOGGLE MODEL TO EXPLAIN THE BALANCE IN SPLICING SPEED WITH FIDELITY

In the work presented here, I propose and test a 'Toggle' model to explain how splicing maintains fidelity while not compromising on speed. This model builds off of several previously presented observations: (1) there are two mutually exclusive conformations taken up by the RH extension in Prp8; (2) helicases drive irreversible activation steps during splicing; (3) the ribosome changes between open and closed conformations to balance speed and fidelity; and (4) group II introns show catalytically silent and active states suggesting that the spliceosome is likely to have two such states as well.

According to the Toggle model, the spliceosome, like the group II intron, also adopts two conformations: a *catalytic* conformation, and a *transitional*, catalytically silent conformation. I propose that the spliceosome toggles back and forth between these conformations to balance splicing speed with fidelity. Further, the two conformations of the RH extension of Prp8 are connected with this structural toggle (Figure 1.10). Specifically, the hairpin conformation of the RH extension is associated with the transitional state while the loop conformation is associated with the catalytic state. The RH extension toggles back and forth between these two conformations during the splicing cycle to balance splicing speed with fidelity.

Similar to the ribosome, I show that this balance can be disrupted by increasing the relative stability of one conformation over the other, and in doing so, mutations in the RH extension that increase the relative stability of the hairpin conformation stabilize the transitional state and show slow, but hyper accurate splicing. On the contrary, mutations in the RH extension that increase the relative stability of the loop conformation stabilize the

catalytic state and show fast, but error-prone splicing. My work proposing and testing this model is described in Chapter 2.

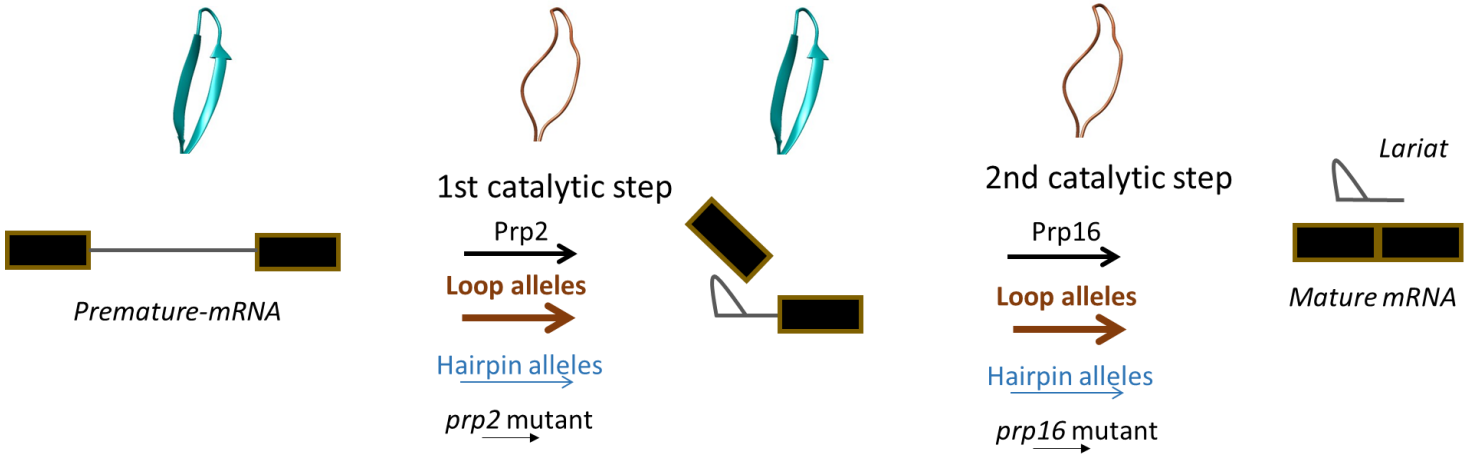


Figure 1.10 Toggle Model

Hairpin (blue structure) conformation of RH extension is associated with the transitional states of the spliceosome before the two steps, while the loop conformation (brown structure) is associated with the catalytic state at both steps. RH extension toggles back and forth between the two conformations to balance splicing speed and fidelity. Mutations that increase the relative stability of the loop conformation stabilize the catalytic state at both steps and splicing is faster, but more error-prone. In contrast, mutations that increase the relative stability of the hairpin conformation stabilize the transitional state at both steps, and splicing is slower, but more accurate.

RH EXTENSION AND SPLICING NATURE OF AN ORGANISM

Given the opposite phenotypes on fidelity/speed associated with stabilizing the hairpin and loop conformations of the RH extension, it is intriguing to consider how the extension has evolved across organisms whose genome-wide substrates vary in their splice site usage. According to the toggle model, the expectation is that, in organisms using rigid splice sites, such as budding yeast, the RH extension would be more stabilized towards the hairpin conformation, yielding a more transitional state with lower speed but higher selectivity. In contrast, organisms that have more flexible splice site selection like fission yeast, flies or humans, the RH extension would be more stabilized towards the loop conformation, yielding a more catalytic/activated state that would accommodate the wider variety of splice site sequences. My work examining this possibility is described in Chapter 3.

REFERENCES

Abelson, J. (2008). Is the spliceosome a ribonucleoprotein enzyme? *Nature Structural & Molecular Biology*, *15*(12), 1235–1237. https://doi.org/10.1038/nsmb1208-1235

Absmeier, E., Santos, K. F., & Wahl, M. C. (2016). Functions and regulation of the Brr2 RNA helicase during splicing. *Cell Cycle (Georgetown, Tex.)*, *15*(24), 3362–3377. https://doi.org/10.1080/15384101.2016.1249549

Collins, C. A., & Guthrie, C. (1999). Allele-specific genetic interactions between Prp8 and RNA active site residues suggest a function for Prp8 at the catalytic core of the spliceosome. *Genes & Development*, *13*(15), 1970–1982. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10444595

Cordin, O., & Beggs, J. D. (2013). RNA helicases in splicing. *RNA Biology*, *10*(1), 83–95. https://doi.org/10.4161/rna.22547

De, I., Schmitzová, J., & Pena, V. (2016). The organization and contribution of helicases to RNA splicing. *Wiley Interdisciplinary Reviews: RNA*, *7*(2), 259–274. https://doi.org/10.1002/wrna.1331

Fica, S. M., & Nagai, K. (2017a). Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nature Structural & Molecular Biology*, *24*(10), 791–799. https://doi.org/10.1038/nsmb.3463

Fica, S. M., & Nagai, K. (2017b). Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nature Structural &*

Molecular Biology, *24*(10), 791–799. https://doi.org/10.1038/nsmb.3463
Fica, S. M., Oubridge, C., Galej, W. P., Wilkinson, M. E., Bai, X.-C., Newman, A. J., & Nagai, K. (2017). Structure of a spliceosome remodelled for exon ligation. *Nature*, *542*(7641), 377–380. https://doi.org/10.1038/nature21078
Fica, S. M., Tuttle, N., Novak, T., Li, N.-S., Lu, J., Koodathingal, P., … Piccirilli, J. A. (2013). RNA catalyses nuclear pre-mRNA splicing. https://doi.org/10.1038/nature12734
Frank, D., Patterson, B., & Guthrie, C. (1992). Synthetic lethal mutations suggest interactions between U5 small nuclear RNA and four proteins required for the second step of splicing. *Molecular and Cellular Biology*, *12*(11), 5197–5205. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1406691
Galej, W. P., Oubridge, C., Newman, A. J., & Nagai, K. (2013). Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature*, *493*(7434), 638–643. https://doi.org/10.1038/nature11843
Galej, W. P., Toor, N., Newman, A. J., & Nagai, K. (2018). Molecular Mechanism and Evolution of Nuclear Pre-mRNA and Group II Intron Splicing: Insights from Cryo-Electron Microscopy Structures. *Chemical Reviews*, acs.chemrev.7b00499. https://doi.org/10.1021/acs.chemrev.7b00499
Galej, W. P., Wilkinson, M. E., Fica, S. M., Oubridge, C., Newman, A. J., & Nagai, K. (2016). Cryo-EM structure of the spliceosome immediately after branching. *Nature*, *537*(7619), 197–201. https://doi.org/10.1038/nature19316
Grainger, R. J., & Beggs, J. D. (2005). Prp8 protein: at the heart of the spliceosome. *RNA (New York, N.Y.)*, *11*(5), 533–557. https://doi.org/10.1261/rna.2220705
Guthrie, C., & Collins, C. A. (2000). The question remains: Is the spliceosome a ribozyme? *Nature Structural Biology*, *7*(10), 850–854. https://doi.org/10.1038/79598
Hotz, H. R., & Schwer, B. (1998). Mutational analysis of the yeast DEAH-box splicing factor Prp16. *Genetics*, *149*(2), 807–815. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9611193
Koodathingal, P., & Staley, J. P. (2013). Splicing fidelity. *RNA Biology*, *10*(7), 1073–1079. https://doi.org/10.4161/rna.25245
Kuhn, A. N., & Brow, D. A. (2000). Suppressors of a cold-sensitive mutation in yeast U4 RNA define five domains in the splicing factor Prp8 that influence spliceosome activation. *Genetics*, *155*(4), 1667–1682. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10924465
Kuhn, A. N., Reichl, E. M., & Brow, D. A. (2002). Distinct domains of splicing factor Prp8 mediate different aspects of spliceosome activation. *Proceedings of the National Academy of Sciences*, *99*(14), 9145–9149. https://doi.org/10.1073/pnas.102304299
Lambowitz, A. M., & Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harbor Perspectives in Biology*, *3*(8), a003616. https://doi.org/10.1101/cshperspect.a003616
Liu, L., Query, C. C., & Konarska, M. M. (2007). Opposing classes of prp8 alleles modulate the transition between the catalytic steps of pre-mRNA splicing. *Nature Structural & Molecular Biology*, *14*(6), 519–526. https://doi.org/10.1038/nsmb1240
Liu, Y.-C., & Cheng, S.-C. (2015). Functional roles of DExD/H-box RNA helicases in Pre-mRNA splicing. *Journal of Biomedical Science*, *22*(1), 54. https://doi.org/10.1186/s12929-015-0161-z
Marcia, M., & Pyle, A. M. (2012a). Visualizing group II intron catalysis through the stages of

splicing. *Cell*. https://doi.org/10.1016/j.cell.2012.09.033
Marcia, M., & Pyle, A. M. (2012b). Visualizing group II intron catalysis through the stages of splicing. *Cell*, *151*(3), 497–507. https://doi.org/10.1016/j.cell.2012.09.033
Martinho, R. G., Guilgur, L. G., & Prudêncio, P. (2015). How gene expression in fast-proliferating cells keeps pace. *BioEssays*, *37*(5), 514–524. https://doi.org/10.1002/bies.201400195
Mayerle, M., Raghavan, M., Ledoux, S., Price, A., Stepankiw, N., Hadjivassiliou, H., … Abelson, J. (2017). Structural toggle in the RNaseH domain of Prp8 helps balance splicing fidelity and catalytic efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(18), 4739–4744. https://doi.org/10.1073/pnas.1701462114
Mozaffari-Jovin, S., Santos, K. F., Hsiao, H.-H., Will, C. L., Urlaub, H., Wahl, M. C., & Lührmann, R. (2012). The Prp8 RNase H-like domain inhibits Brr2-mediated U4/U6 snRNA unwinding by blocking Brr2 loading onto the U4 snRNA. *Genes & Development*, *26*(21), 2422–2434. https://doi.org/10.1101/gad.200949.112
Ogle, J. M., Murphy, F. V., Tarry, M. J., & Ramakrishnan, V. (2002). Selection of tRNA by the Ribosome Requires a Transition from an Open to a Closed Form. *Cell*, *111*(5), 721–732. https://doi.org/10.1016/S0092-8674(02)01086-3
Ogle, J. M., & Ramakrishnan, V. (2005). STRUCTURAL INSIGHTS INTO TRANSLATIONAL FIDELITY. *Annual Review of Biochemistry*, *74*(1), 129–177. https://doi.org/10.1146/annurev.biochem.74.061903.155440
Pena, V., Rozov, A., Fabrizio, P., Lührmann, R., & Wahl, M. C. (2008). Structure and function of an RNase H domain at the heart of the spliceosome. *The EMBO Journal*, *27*(21), 2929–2940. https://doi.org/10.1038/emboj.2008.209
Plaschka, C., Lin, P.-C., & Nagai, K. (2017). Structure of a pre-catalytic spliceosome. *Nature*, *546*(7660), 617–621. https://doi.org/10.1038/nature22799
Pyle, A. M. (2016). Group II Intron Self-Splicing. *Annual Review of Biophysics*, *45*(1), 183–205. https://doi.org/10.1146/annurev-biophys-062215-011149
Query, C. C., & Konarska, M. M. (2004a). Suppression of Multiple Substrate Mutations by Spliceosomal prp8 Alleles Suggests Functional Correlations with Ribosomal Ambiguity Mutants. *Molecular Cell*, *14*(3), 343–354. https://doi.org/10.1016/S1097-2765(04)00217-5
Query, C. C., & Konarska, M. M. (2004b). Suppression of Multiple Substrate Mutations by Spliceosomal prp8 Alleles Suggests Functional Correlations with Ribosomal Ambiguity Mutants. *Molecular Cell*, *14*(3), 343–354. https://doi.org/10.1016/S1097-2765(04)00217-5
Query, C. C., & Konarska, M. M. (2006). Splicing fidelity revisited. *Nature Structural & Molecular Biology*, *13*(6), 472–474. https://doi.org/10.1038/nsmb0606-472
Rauhut, R., Fabrizio, P., Dybkov, O., Hartmuth, K., Pena, V., Chari, A., … Luhrmann, R. (2016). Molecular architecture of the Saccharomyces cerevisiae activated spliceosome. *Science*, *353*(6306), 1399–1405. https://doi.org/10.1126/science.aag1906
Ritchie, D. B., Schellenberg, M. J., Gesner, E. M., Raithatha, S. A., Stuart, D. T., & MacMillan, A. M. (2008). Structural elucidation of a PRP8 core domain from the heart of the spliceosome. *Nature Structural & Molecular Biology*, *15*(11), 1199–1205. https://doi.org/10.1038/nsmb.1505
Rodnina, M. V, Fischer, N., Maracci, C., & Stark, H. (2017). Ribosome dynamics during decoding. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1716). https://doi.org/10.1098/rstb.2016.0182
Schellenberg, M. J., Wu, T., Ritchie, D. B., Fica, S., Staley, J. P., Atta, K. A., … MacMillan, A.

M. (2013a). A conformational switch in PRP8 mediates metal ion coordination that promotes pre-mRNA exon ligation. *Nature Structural & Molecular Biology*, *20*(6), 728–734. https://doi.org/10.1038/nsmb.2556

Schellenberg, M. J., Wu, T., Ritchie, D. B., Fica, S., Staley, J. P., Atta, K. A., … MacMillan, A. M. (2013b). A conformational switch in PRP8 mediates metal ion coordination that promotes pre-mRNA exon ligation. *Nature Structural & Molecular Biology*, *20*(6), 728–734. https://doi.org/10.1038/nsmb.2556

Scheres, S. H., & Nagai, K. (2017a). CryoEM structures of spliceosomal complexes reveal the molecular mechanism of pre-mRNA splicing. *Current Opinion in Structural Biology*, *46*, 130–139. https://doi.org/10.1016/j.sbi.2017.08.001

Scheres, S. H., & Nagai, K. (2017b). CryoEM structures of spliceosomal complexes reveal the molecular mechanism of pre-mRNA splicing. *Current Opinion in Structural Biology*, *46*, 130–139. https://doi.org/10.1016/J.SBI.2017.08.001

Schneider, S., Campodonico, E., & Schwer, B. (2004). Motifs IV and V in the DEAH Box Splicing Factor Prp22 Are Important for RNA Unwinding, and Helicase-defective Prp22 Mutants Are Suppressed by Prp8. *Journal of Biological Chemistry*, *279*(10), 8617–8626. https://doi.org/10.1074/jbc.M312715200

Sharp, P. A. (2005). The discovery of split genes and RNA splicing. *Trends in Biochemical Sciences*, *30*(6), 279–281. https://doi.org/10.1016/j.tibs.2005.04.002

Shi, Y. (2017). The Spliceosome: A Protein-Directed Metalloribozyme. *Journal of Molecular Biology*, *429*(17), 2640–2653. https://doi.org/10.1016/J.JMB.2017.07.010

Siatecka, M., Reyes, J. L., & Konarska, M. M. (1999). Functional interactions of Prp8 with both splice sites at the spliceosomal catalytic center. *Genes & Development*, *13*(15), 1983–1993. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10444596

Singh, R. K., & Cooper, T. A. (2012). Pre-mRNA splicing in disease and therapeutics. *Trends in Molecular Medicine*, *18*(8), 472–482. https://doi.org/10.1016/j.molmed.2012.06.006

Steitz, T. A., & Steitz, J. A. (1993). A general two-metal-ion mechanism for catalytic RNA. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(14), 6498–6502. https://doi.org/10.1073/pnas.90.14.6498

Toor, N., Keating, K. S., Taylor, S. D., & Pyle, A. M. (2008). Crystal structure of a self-spliced group II intron. *Science (New York, N.Y.)*, *320*(5872), 77–82. https://doi.org/10.1126/science.1153803

Umen, J. G., & Guthrie, C. (n.d.). Mutagenesis of the Yeast Gene P W 8 Reveals Domains Governing the Specificity and Fidelity of 3’ Splice Site Selection. Retrieved from http://www.genetics.org/content/genetics/143/2/723.full.pdf

Umen, J. G., & Guthrie, C. (1995). Prp16p, Slu7p, and Prp8p interact with the 3’ splice site in two distinct stages during the second catalytic step of pre-mRNA splicing. *RNA (New York, N.Y.)*, *1*(6), 584–597. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7489518

Wahl, M. C., Will, C. L., & Lührmann, R. (2009a). The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell*. https://doi.org/10.1016/j.cell.2009.02.009

Wahl, M. C., Will, C. L., & Lührmann, R. (2009b). The spliceosome: design principles of a dynamic RNP machine. *Cell*, *136*(4), 701–718. https://doi.org/10.1016/j.cell.2009.02.009

Wan, R., Yan, C., Bai, R., Huang, G., & Shi, Y. (2016). Structure of a yeast catalytic step I spliceosome at 3.4 A resolution. *Science*, *353*(6302), 895–904. https://doi.org/10.1126/science.aag2235

Will, C. L., & Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor*

Perspectives in Biology, *3*(7). https://doi.org/10.1101/cshperspect.a003707
Xu, D., Field, D. J., Tang, S. J., Moris, A., Bobechko, B. P., & Friesen, J. D. (1998). Synthetic lethality of yeast slt mutations with U2 small nuclear RNA mutations suggests functional interactions between U2 and U5 snRNPs that are important for both steps of pre-mRNA splicing. *Molecular and Cellular Biology*, *18*(4), 2055–2066. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9528778
Yan, C., Wan, R., Bai, R., Huang, G., & Shi, Y. (2016). Structure of a yeast activated spliceosome at 3.5 A resolution. *Science*, *353*(6302), 904–911. https://doi.org/10.1126/science.aag0291
Yan, C., Wan, R., Bai, R., Huang, G., & Shi, Y. (2017). Structure of a yeast step II catalytically activated spliceosome. *Science*, *355*(6321), 149–155. https://doi.org/10.1126/science.aak9979
Yang, K., Zhang, L., Xu, T., Heroux, A., & Zhao, R. (2008). Crystal structure of the beta-finger domain of Prp8 reveals analogy to ribosomal proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(37), 13817–13822. https://doi.org/10.1073/pnas.0805960105

CHAPTER 2

STRUCTURAL TOGGLE IN THE RNASEH DOMAIN OF PRP8 HELPS BALANCE SPLICING FIDELITY AND CATALYTIC EFFICIENCY

Megan Mayerle,[a,1] Madhura Raghavan,[b,1] Sarah Ledoux,[a,1] Argenta Price,[a,1,2] Nicholas Stepankiw,[b] Haralambos Hadjivassiliou,[a] Erica A. Moehle,[a,3] Senén D. Mendoza,[a] Jeffrey A. Pleiss,[b,4] Christine Guthrie,[a,4] and John Abelson[a,4]

Author contributions: M.M., M.R., S.L., A.P., N.S., H.H., J.A.P., C.G., and J.A. designed research; M.M., M.R., S.L., A.P., N.S., H.H., E.A.M., and S.D.M. performed research; M.M., M.R., S.L., A.P., N.S., and H.H. contributed new reagents/analytic tools; M.M., M.R., S.L., A.P., N.S., S.D.M., J.A.P., C.G., and J.A. analyzed data; and M.M., M.R., S.L., A.P., J.A.P., C.G., and J.A. wrote the paper.

[1]M.M., M.R., S.L., and A.P. contributed equally to this work.
[2]Present address: Department of Physics, University of Colorado, Boulder, CO 80309.
[3]Present address: Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720.
[a]Department of Biochemistry and Biophysics, University of California, San Francisco, CA.
[b]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY.

This work was a collaboration with the Guthrie & Abelson group at UCSF. I was involved in the part investigating genome-wide effects on splicing fidelity using lariat sequencing and RT-PCR. This work was published in the journal, Proceedings of the National Academy of Sciences.

Mayerle, M., Raghavan, M., Ledoux, S., Price, A., Stepankiw, N., Hadjivassiliou, H., Moehle, E., Mendoza, S.D., Pleiss, J.A., Guthrie, C. & Abelson, J. (2017). Structural toggle in the RNaseH domain of Prp8 helps balance splicing fidelity and catalytic efficiency. Proceedings of the National Academy of Sciences of the United States of America, 114(18), 4739–4744.

ABSTRACT

Pre-mRNA splicing is an essential step of eukaryotic gene expression that requires both high efficiency and high fidelity. Prp8 has long been considered the "master regulator" of the spliceosome, the molecular machine that executes pre-mRNA splicing. Cross-linking and structural studies place the RNaseH domain (RH) of Prp8 near the spliceosome's catalytic core and demonstrate that *prp8* alleles that map to a 17-aa extension in RH stabilize it in one of two mutually exclusive structures, the biological relevance of which are unknown. We performed an extensive characterization of *prp8* alleles that map to this extension and, using in vitro and in vivo reporter assays, show they fall into two functional classes associated with the two structures: those that promote error-prone/efficient splicing and those that promote hyperaccurate/inefficient splicing. Identification of global locations of endogenous splice-site activation by lariat sequencing confirms the fidelity effects seen in our reporter assays. Furthermore, we show that error-prone/efficient RH alleles suppress a *prp2* mutant deficient at promoting the first catalytic step of splicing, whereas hyperaccurate/inefficient RH alleles exhibit synthetic sickness. Together our data indicate that *prp8* RH alleles link splicing fidelity with catalytic efficiency by biasing the relative stabilities of distinct spliceosome conformations. We hypothesize that the spliceosome "toggles" between such error-prone/efficient and hyperaccurate/inefficient conformations during the splicing cycle to regulate splicing fidelity.

INTRODUCTION

Pre-mRNA splicing occurs via two transesterification reactions catalyzed by the spliceosome, a large, dynamic ribonucleoprotein complex. The catalytically active spliceosome is composed of three small nuclear ribonucleoprotein (snRNP) complexes (U2, U5, and U6), the Nineteen Complex and its related proteins, and a small number of accessory splicing proteins (1). In the first catalytic step of pre-mRNA splicing the 5′ splice site (5′SS) is cleaved, forming a lariat-3′ exon intermediate in which the 5′SS is covalently linked via a 2′,5′ phosphodiester bond to the branch site adenosine (BrA). In the second catalytic step the 3′ splice site (3′SS) is cleaved and the 5′ and 3′ exons are ligated. The excised lariat intron is then released from the spliceosome. Splicing is performed with high fidelity and high efficiency, and it is thought that core spliceosomal components contribute directly to maintenance of splicing fidelity and efficiency (1, 2).

Prp8 is the largest and most highly conserved protein in the spliceosome and is a component of the U5 snRNP. Recent cryoelectron microscopy structures show that Prp8 acts as a platform at the heart of the spliceosome and undergoes considerable conformational rearrangements during the splicing cycle (3–11). Genetic screens have identified many alleles in Prp8 that exhibit compromised splicing fidelity or efficiency (summarized in ref. 12). A subset of these mutations maps to a unique and essential 17-aa extension within the RNaseH domain (RH) of Prp8 (13–20). Structural studies reveal that the RH extension can exist either in a β-hairpin form or as a disordered loop, and that it adopts these conformations at distinct steps of the splicing cycle (Figure 2.1) (3, 4, 7–11, 19). Of the subset of *prp8* alleles that map to the RH extension for which structural data are available,

those that preferentially stabilize either the β-hairpin or disordered loop forms of RH correlate with distinct genetic phenotypes previously proposed to arise from unique first-step and second-step catalytic conformations of the spliceosome (Figure 2.1 and Table 2.1) (17–20).

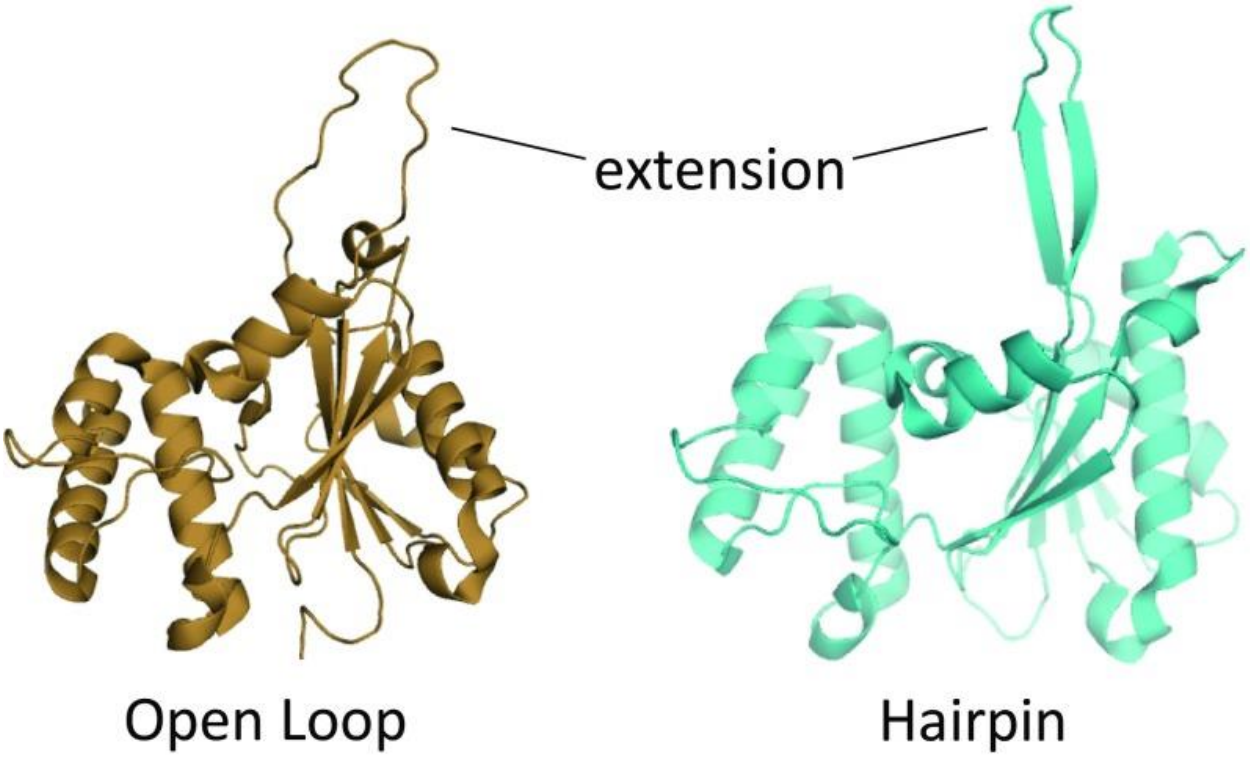


Figure 2.1 The 17-aa extension in the RNaseH subdomain of Prp8 adopts two different forms: an open loop (tan, *Left*) and a β-hairpin (cyan, *Right*). Structures are modified from Protein Data Bank ID code 4JK7 (19).

Table 2.1

Summary of previous data and characterization of *prp8* toggle alleles

Yeast/ (human)	Refs.	Original identification	Toggle classification	Information from human RH structure	Structure favored	Growth
V1860D (V1788)	13, 14	U4-cs1 suppressor	Transitional/first	Structure solved: H-bond with Y1786 stabilizes hairpin	Hairpin (19)	As WT
V1860N (V1788)	14	U4-cs1 suppressor	Transitional	Predicted H-bond with Y1786	Hairpin (our prediction)	As WT
V1862D (I1790)	13, 14	U4-cs1 suppressor	Transitional	Predicted H-bond with T1800	Hairpin (our prediction)	As WT
T1865K (T1793)	13	Designed	Transitional/first	Predicted H-bond with N1797	Hairpin (our prediction)	As WT
A1871E (T1799)	13	Designed	Transitional/first	Predicted H-bond with H1791	Hairpin (our prediction)	As WT
T1872E (T1800)	13	Designed	Transitional/first	Structure solved: H2O-mediated H-bond with Y1786 stabilizes hairpin	Hairpin (19)	As WT
T1861P (T1789)	15, 19	U4-cs1 suppressor	Catalytic/second	Structure solved: proline disrupts hairpin	Loop (19)	As WT
V1862Y (I1790)	14, 19	U4-cs1 suppressor	Catalytic/second	Structure solved: H-bond with N1797 stabilizes loop	Loop (19)	As WT
H1863E (H1791)	13	Designed	Catalytic/second	No prediction	Unknown	As WT

K1864E (K1792)	16	5′SS suppressor	Catalytic/second	No prediction	Unknown	As WT
N1869D (N1797)	16, 17, 20	5′SS suppressor	Catalytic/second	No prediction	Unknown	Ts (slightly)
V1870N (L1798)	17, 18	BrG suppressor	Catalytic/second	Predicted H-bond with G1796	Loop (19)	As WT
I1875T (I1803)	15	U4-cs1 suppressor	Catalytic	Predicted H-bond with R1787	Loop (our prediction)	As WT
Δ1860–1875	This study	Designed	unclassified	Removes hairpin/loop	——	Lethal

However, recent structural and biochemical studies of the spliceosome and its evolutionary precursor, the group II intron, reveal that the spliceosome's catalytic core is similar at both catalytic steps (21, 22). Structural data further suggest the group II intron passes through an obligatory intermediate transitional structure between two highly similar first- and second-step catalytic structures (22). Conservation argues that some components of the spliceosome might adopt a similar transitional intermediate, "toggling" between catalytic and transitional intermediate conformations during a typical splicing reaction, and that these conformations would be executed through specific structural toggles in core spliceosomal components (17).

Here we provide evidence for two distinct classes of *prp8* RH alleles, which we refer to as "catalytic" and "transitional" in keeping with the nomenclature established for the group II intron. Specifically, catalytic *prp8* alleles that stabilize the loop structure of the RH extension exhibit high-efficiency, low-fidelity splicing, whereas transitional *prp8* alleles that stabilize the β-hairpin structure exhibit low-efficiency but high-fidelity splicing both on reporter

constructs and at endogenous splice sites genomewide. We propose that the spliceosome cycles between catalytic and transitional conformations during each splicing cycle, and we implicate the RH extension of Prp8 as a structural toggle. This model is supported by recent high-resolution structural studies of the spliceosome that show that the RH extension adopts a loop form in the catalytically active B^{act} spliceosome (5) and a β-hairpin form in other noncatalytically active conformations (3, 4, 7–9). Our data provide a specific, mechanistic example of how distinct conformations of a core spliceosomal component can affect both splicing fidelity and efficiency.

RESULTS

Creation of *prp8* Toggle Allele Strains

We chose 13 published *prp8* RH extension alleles for characterization: V1860D, V1860N, T1861P, V1862D, V1862Y, H1863E, K1864E, T1865K, N1869D, V1870N, A1871E, T1872E, and I1875T. Together, we refer to these alleles as the "toggle" alleles. Whereas most of these alleles were discovered through genetic screens conducted to identify alleles that suppress growth defects caused by the mutation of the 5′SS, branch point (Br), or 3′SS motifs of a reporter gene (16–18, 20, 23), a subset was deliberately designed based on structural data (13) (Table 2.1). Structural data are available for a subset of these alleles (19); for the other alleles, we made predictions based on the available data (Table 2.1). Most alleles grew similarly to WT yeast in rich media at all temperatures tested, consistent with previous results. A strain in which the entire 17-aa RH extension was deleted failed to grow at any temperature (Table 2.1).

ACT1-CUP1 Reporter Assay Demonstrates That *prp8* Toggle Alleles Sort into Two Distinct Classes

To determine the effects of the *prp8* toggle alleles on splicing fidelity and efficiency we used the well-characterized *ACT1-CUP1* splicing reporter system (Figure 2.2*A* and Table 2.2) (24). In this system, the ability of yeast to grow on otherwise lethal concentrations of copper-containing media is directly proportional to splicing efficiency. *ACT1-CUP1* reporters containing nonconsensus splicing sequences at the 5′SS (G1A, U2A, and G5A), Br (C256A, BrC, and BrG), or 3′SS (U301G, A302G, and G303/304C) were used to

measure fidelity (Figure 2.2 *B* and *C*). A false-color representation of the data is shown in Figure 2.2*D* (25), and the results are consistent with reported data for the subset of toggle alleles previously examined (Figure 2.3) (13, 16–20).

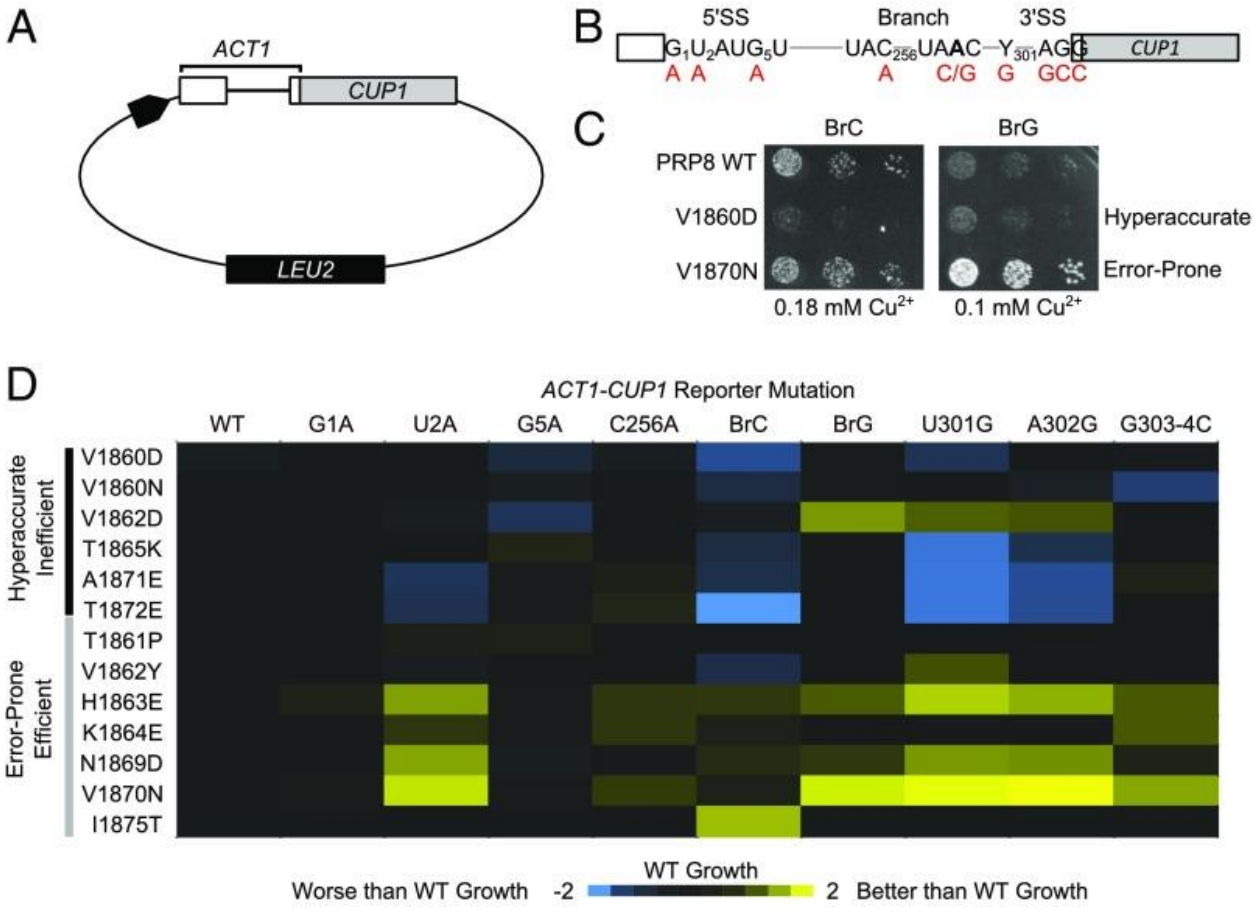


Figure 2.2 (A) Schematic of ACT1-CUP1 reporter. (B) Diagram of ACT1-CUP1 reporter intron. The 5′SS, BrA, and 3′SS are shown with mutations made to test fidelity in red. (C) Growth of prp8 toggle alleles in the presence of BrC (Left) or BrG (Right). (D) [Cu2+max] that supports growth was determined for each prp8 toggle allele and reporter. Values were transformed log2([Cu2+max prp8]/[Cu2+max PRP8]) and colored blue (worse growth) to yellow (better growth).

Table 2.2

Effects of *ACT1-CUP1* reporters used as shown previously by primer extension analysis (17, 18, 24). Numbering is based on the ACT1-CUP1 reporter described in (24).

Reporter	Effect	Refs.
G1A	Inhibits second step	24
U2A	Inhibits both steps, lariat intermediate degraded	18
G5A	Inhibits both steps, activates aberrant 5′SS use	24
C256A	Inhibits both steps	24
BrC	Inhibits both steps, limiting for first step	17, 18
BrG	Inhibits second step	17
U301G (gAG)	Inhibits second step, lariat intermediate degraded	18
A302G (UgG)	Inhibits second step, lariat intermediate degraded	18
G303/4C (UAc/c)	Inhibits second step	17

	G1A	U2A	U2G	A3C	G5A	C256A	BrC	BrG	U301G	A302G	A302U	G303/4C
V1860D	g	g			g	g	g	f,g	g	g		g
V1860N	g	g			g	g	g	g	g	g		g
V1862D	g	g			g	g	g	e,g	g	g	e	g
T1865K	g	g			g	g	g	c,g	g	g	e	g
A1871E	g	g			g	g	g	e,g	g	g		g
T1872E	g	g			g	g	g	e,f,g	g	g	e	g
T1861P	g	g			g	g	g	f,g	g	g		g
V1862Y	g	g			g	g	g	f,g	g	g		g
H1863E	g	g			g	g	g	e,g	g	g	e	g
K1864E	g	c,g			g	g	c,g	e,g	g	g	c	g
N1869D	a,b,c,g	a,b,c,g			b,g	c,g	b,c,g	c,g	a,b,c,g	g	a,b,c	b,c,g
V1870N	g	c,g	a,b,c	d	g	g	c,d,g	c,d,e,f,g	g	g	c,d,e	g
I1875T	g	g			g	g	g	g	g	g		g

Figure 2.3 Comparison of ACT1-CUP1 growth assays. Data from seven different studies (13, 17–20, 42) are included with the data generated in this work. Background coloring indicates growth of prp8 toggle allele carrying the indicated ACT1-CUP1 splicing reporter, where black indicates WT growth, blue indicates decreased/hyperaccurate growth in at least one reference, yellow indicates increased/error-prone growth in at least one reference, and gray untested. References used to create this table are indicated by letter: a, ref. 20; b, ref. 16; c, ref. 17; d, ref. 18; e, ref. 13; f, ref. 19; and g, this work. Red letter coloring indicates that in the indicated reference the prp8 allele grew as WT. Such differences are likely due to differing strain backgrounds used by different laboratories.

All *prp8* toggle alleles grew with efficiency similar to *PRP8* when required to splice a WT *ACT1-CUP1*reporter. However, individual *prp8* toggle alleles exhibited differential growth on *ACT1-CUP1* reporters with mutated, nonconsensus splicing sequences (Figure 2.2*D* and Figure 2.3). In general, *prp8* toggle alleles sorted into two groups: those that exhibited worse growth (blue, Figure 2.2*D*) than *PRP8* (indicative of hyperaccurate/inefficient splicing) and those that exhibited better growth (indicative of error-prone splicing) (yellow, Figure 2.2*D*). This division held whether the reporter affected primarily the first step of splicing, the second step, or both (Table 2.2). Differences in the

extent of the effect varied with individual *prp8* toggle alleles. For example, V1860D and V1860N both exhibited a hyperaccurate/inefficient phenotype; however, the phenotype was much stronger in *prp8* V1860D regardless of reporter. Such variability is not unexpected, because toggle alleles may bias RH domain conformation to differing degrees. These data suggested that *prp8* toggle alleles could be broadly classified as exhibiting hyperaccurate/inefficient or error-prone/efficient splicing, associated with the β-hairpin and loop conformations of the RH extension, respectively.

Toggle Alleles Interact Genetically with an ATPase-Deficient *prp2* Allele

The DEAH-box helicase Prp2 is required for the first catalytic step of splicing (1). Prp2 destabilizes the association between the U2 snRNP and the pre-mRNA before the first step (26) and promotes snRNA rearrangements in preparation for catalysis (27). Strains containing the *prp2*-Q548E allele are impaired for growth at 16 °C and exhibit defects in pre-mRNA splicing, likely due to deficiencies in ATP binding and/or hydrolysis that lead to inefficient catalytic activation (27). We made double-mutant strains that contained *prp2*-Q548E in combination with *prp8* alleles from each class identified in our *ACT1-CUP1* reporter assay: V1860D and V1860N (hyperaccurate/inefficient) and N1869D and V1870N (error-prone/efficient). Hyperaccurate/inefficient *prp8* alleles exhibited synthetic sickness at 16 °C when combined with *prp2*-Q548E whereas error-prone/efficient *prp8* toggle alleles rescued the cold sensitivity of the *prp2*-Q548E strain (Figure 2.4). Because Prp2 is required for catalytic activation, these data are consistent with a model in which error-prone/efficient *prp8* alleles promote splicing catalysis whereas hyperaccurate/inefficient oppose it.

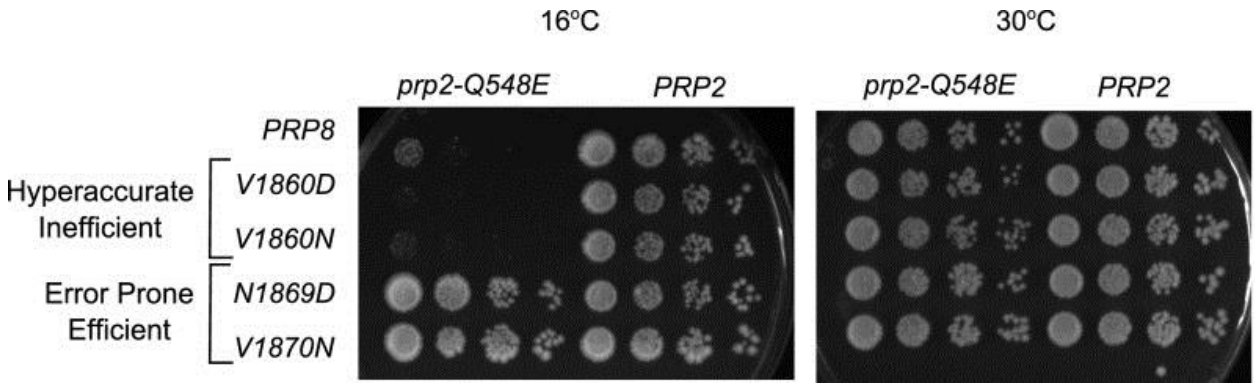


Figure 2.4 Growth of double-mutant strains carrying WT, V1860D, V1860N, N1869D, and V1870N *prp8* alleles in combination with *PRP2* or *prp2-Q548E* at 16 °C (*Left*) and 30 °C (*Right*).

In Vitro Characterization of the First-Step Splicing Efficiency of *prp8* Toggle Mutants

To more directly assess first-step splicing efficiency we performed in vitro splicing assays using a splicing substrate truncated just downstream of the branch site. This substrate is unable to complete the second step because it lacks a 3′SS (19, 28), enabling more robust characterization of the first step. Figure 2.5*A* shows a representative time-course experiment performed in *PRP8* extract and extracts made from two toggle alleles (V1860D and V1870N) that, based upon *ACT1-CUP1* reporter assays, were expected to have opposite effects on splicing efficiency. Whereas extracts made from error-prone/efficient *prp8* toggle alleles spliced the truncated pre-mRNA with efficiency similar to WT, those from hyperaccurate/inefficient *prp8* alleles exhibited decreased efficiency.

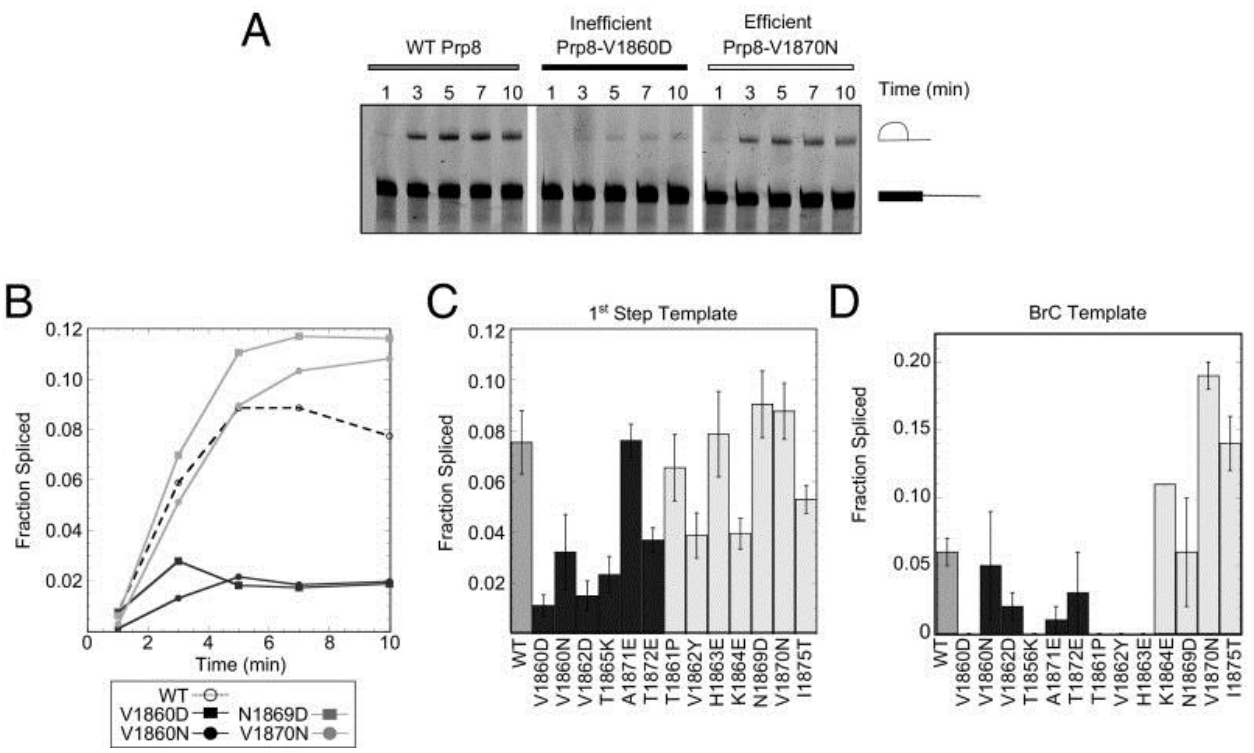


Figure 2.5 (A) First-step in vitro splicing assay. Gel showing results of a time course of splicing of a fluorescent pre-Act1 truncated template. Pre-mRNA and first-step product indicated. (B) Quantification of representative first-step time course assay. (C) Fraction truncated pre-Act1 spliced after 10 min. (D) Fraction pre-BrC template spliced after 20 min is shown. Hyperaccurate/inefficient alleles colored black, error-prone/efficient gray. Error bars are SEM for three biological replicates.

Because the first-step reaction was essentially complete at 10 min (Figure 2.5*B*), we repeated this analysis with extracts made from all *prp8* toggle alleles but focused only on the 10-min time point (Figure 2.5*C*). On the whole, hyperaccurate/inefficient alleles performed the first step of splicing less efficiently than WT, whereas error-prone/efficient alleles performed the first step with similar or increased efficiency. There were a few exceptions: Three of the *prp8* toggle alleles classified as error-prone/efficient based on *ACT1-CUP1* reporter data (V1862Y, K1864E, and I1875T) spliced with lower efficiency, whereas

the hyperaccurate/inefficient *prp8* allele A1871E spliced with efficiency similar to WT. This might reflect allele-specific variability in the artificial context of in vitro splicing with a truncated pre-mRNA and hints at additional complexity in the mode of action of the *prp8* toggle alleles.

To further characterize first-step catalytic efficiency we used extracts made from *prp8* toggle alleles to perform in vitro splicing assays on a full-length *ACT1* pre-mRNA harboring a BrC mutation (Figure 2.5*D* and Figure 2.6). The BrC mutation decreases the efficiency of both steps, with a particularly strong effect on the first step (17, 29). Extracts from most of the error-prone/efficient *prp8* toggle alleles spliced a BrC-containing substrate more efficiently than *PRP8*. This included extracts from two of the alleles that had spliced the truncated pre-mRNA substrate less efficiently, further indicating potential template-specific effects. All extracts made from hyperaccurate/inefficient *prp8* alleles spliced BrC template less efficiently than WT. Some extracts from both *prp8* toggle allele classes were unable to splice mutant templates in vitro. Because all of the extracts spliced the WT template, we presume that this inefficiency reflects a specific defect between the mutant template and these extracts.

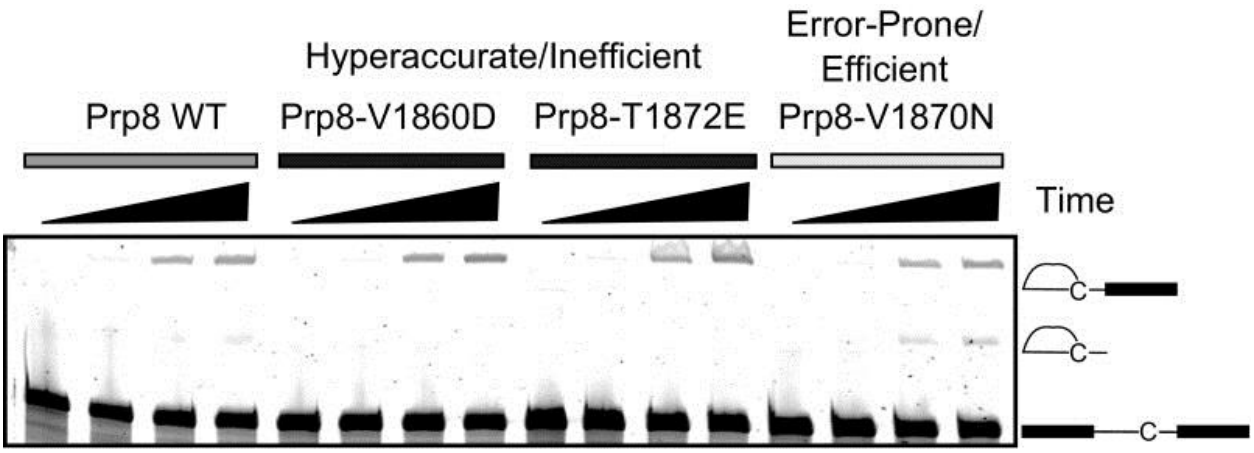


Figure 2.6 Extracts made from error-prone/efficient prp8 alleles perform both steps of splicing more efficiently than those made from WT PRP8, whereas extracts made from hyperaccurate/inefficient prp8 alleles perform as those made from WT PRP8 or worse. Gel showing results of a time-course experiment comparing splicing of a fluorescent pre-Act1 template carrying a BrC mutation. Pre-mRNA, lariat intermediate first-step product and excised lariat second-step product are indicated.

Lariat Sequencing Reveals Genome-wide Alterations to Splicing Fidelity

Given the altered fidelity of *prp8* RH alleles as revealed by reporter constructs, we sought to assess the global impacts of *prp8* RH alleles on the in vivo splice-site selection for native introns. Several groups, including ours, have recently described methods for determining the global locations of splice-site activation by monitoring the lariats generated during each splicing reaction (30–34). Here we developed a modified approach that enables quantitative assessment of splice-site use (*Materials and Methods*). Strains were created carrying the *prp8* toggle alleles in combination with a deletion of the *DBR1* gene, an essential component of the lariat decay pathway, to facilitate lariat accumulation. Lariat RNAs were enriched from the total RNA by first depleting the sample of rRNAs and then enzymatically degrading linear RNAs. When cDNA generated from lariat RNAs is subjected to high-

throughput sequencing, a subset of sequencing reads that traverse the 2′,5′ linkage can be used to extract the specific 5′SS and Br sequences that were activated in the splicing reaction (32). Although reverse transcriptase exhibits a low propensity to transcribe across the 2′,5′ linkage under standard reaction conditions, here we identified conditions that allow for an ~100-fold increase in the frequency of read-through (Figure 2.7).

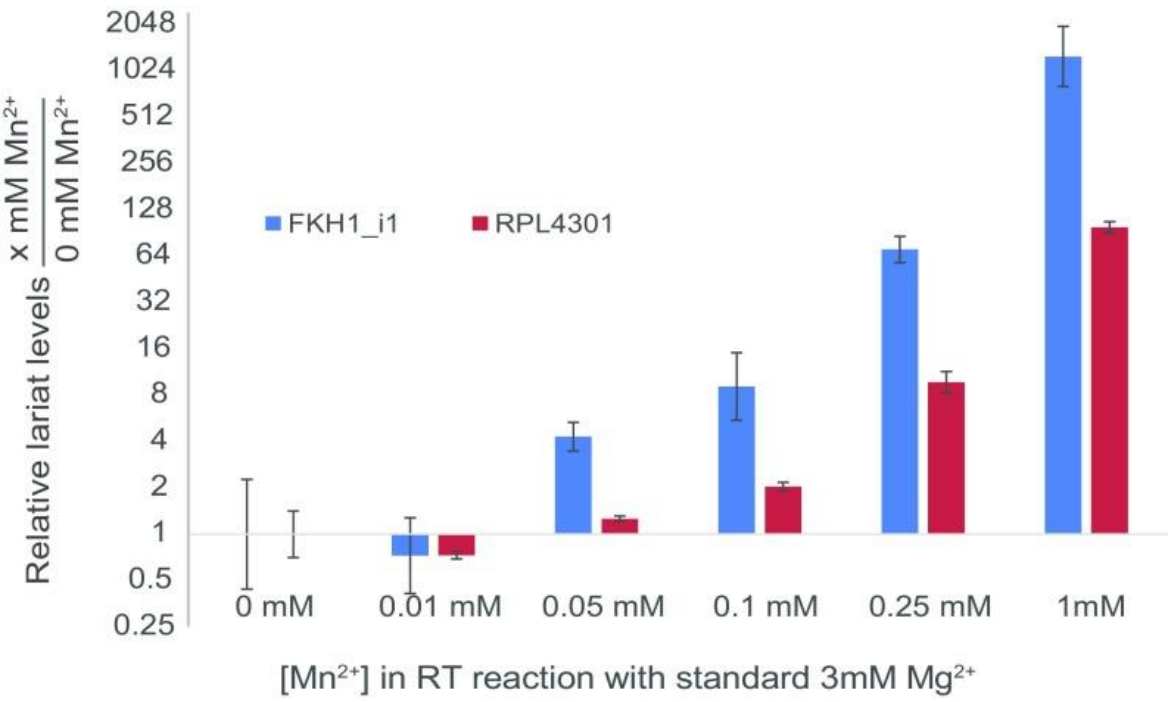
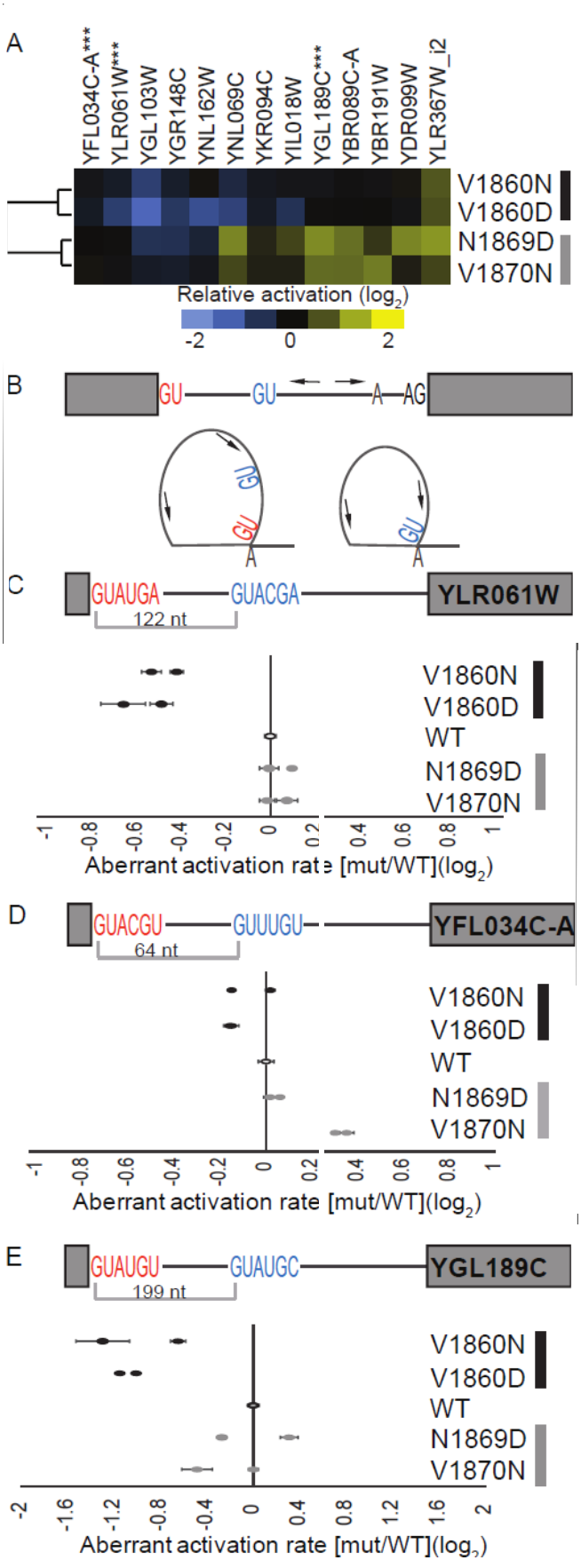


Figure 2.7 Effect of addition of Mn2+ during cDNA synthesis on relative lariat levels are shown for two *Schizosaccharomyces pombe* intron lariats, *fkh1*_Intron1 and *rpl4301*. Confidence intervals are based on three qPCR technical replicates. cDNA synthesis was performed in standard RT buffer containing 3 mM Mg^{2+} along with a titration series of $MnCl_2$. The efficiency of RT reading through the 2′,5′ linkage in the lariat branch was measured as a function of number of lariat amplicons available for amplification in a quantitative PCR assay done on the cDNA with primers that were designed such that they faced away from one another in the linear RNA but toward one another in the context of a lariat, as in Figure 2.8. Relative lariat levels were then calculated by normalization with an exon in the corresponding gene under investigation. Then, using the ΔΔCT method of analysis, fold change in relative lariat levels between each sample and the sample made with 0 mM Mn^{2+} was calculated. Addition of 1 mM Mn^{2+} in the RT reaction resulted in around a 100-fold increase in relative lariat levels over a sample made with no Mn^{2+}. It is to be noted that the exon levels remained largely unchanged as measured by raw CT values, indicating that the increase in relative lariat levels with increasing [Mn^{2+}] was largely because of an increase in lariat amplicon available for amplification as a result of RT reading through the lariat branch.

We identified the global locations of splice-site activation in WT *PRP8*, as well as in two *prp8* toggle alleles from each of the hyperaccurate/inefficient (V1860D and V1860N) and error-prone/efficient (N1869D and V1870N) classes (Figure 2.8A). When considering only the WT *PRP8* strain, lariat reads were detected for 287 of the 308 (93%) annotated spliceosomal introns, a level of coverage that exceeded that obtained by other published lariat sequencing approaches in *Saccharomyces cerevisiae* (33, 34) and established the capacity of this approach to readily capture global sites of splicing activation (Figure 2.9 and Datasets S1 and S2 obtainable from (http://www.pnas.org/content/114/18/4739/tab-figures-data). Consistent with previous studies, this analysis also revealed use of alternative splice sites associated with known introns (30–34), herein referred to as aberrant splice sites. We focused further analyses on the aberrant events with sufficient read depth across all strains to enable statistical analyses.

Figure 2.8 *A*) Global lariat sequencing identified specific locations of increases (yellow) and decreases (blue) in aberrant 5′SS activation associated with annotated introns. (*B*) Schematic of primer locations (arrows) that enable discrimination between annotated (red) and aberrant (blue) splice-site activation. (*C–E*) Relative splice-site use as determined by RT-PCR is shown for two biological replicates of each strain. Hyperaccurate/inefficient alleles colored black, error-prone/efficient gray. Error bars are SD for two technical replicates.


A
YFL034C-A***
YLR061W***
YGL103W
YGR148C
YNL162W
YNL069C
YKR094C
YIL018W
YGL189C***
YBR089C-A
YBR191W
YDR099W
YLR367W_i2
V1860N
V1860D
N1869D
V1870N
Relative activation (log2)
-2
0
2
B
GU
GU
A
AG
GU
GU
A
GU
A
C
GUAUGA
GUACGA
YLR061W
122 nt
V1860N
V1860D
WT
N1869D
V1870N
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Aberrant activation rate [mut/WT](log2)
D
GUACGU
GUUUGU
YFL034C-A
64 nt
V1860N
V1860D
WT
N1869D
V1870N
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Aberrant activation rate [mut/WT](log2)
E
GUAUGU
GUAUGC
YGL189C
199 nt
V1860N
V1860D
WT
N1869D
V1870N
-2
-1.6
-1.2
-0.8
-0.4
0
0.4
0.8
1.2
1.6
2
Aberrant activation rate [mut/WT](log2)

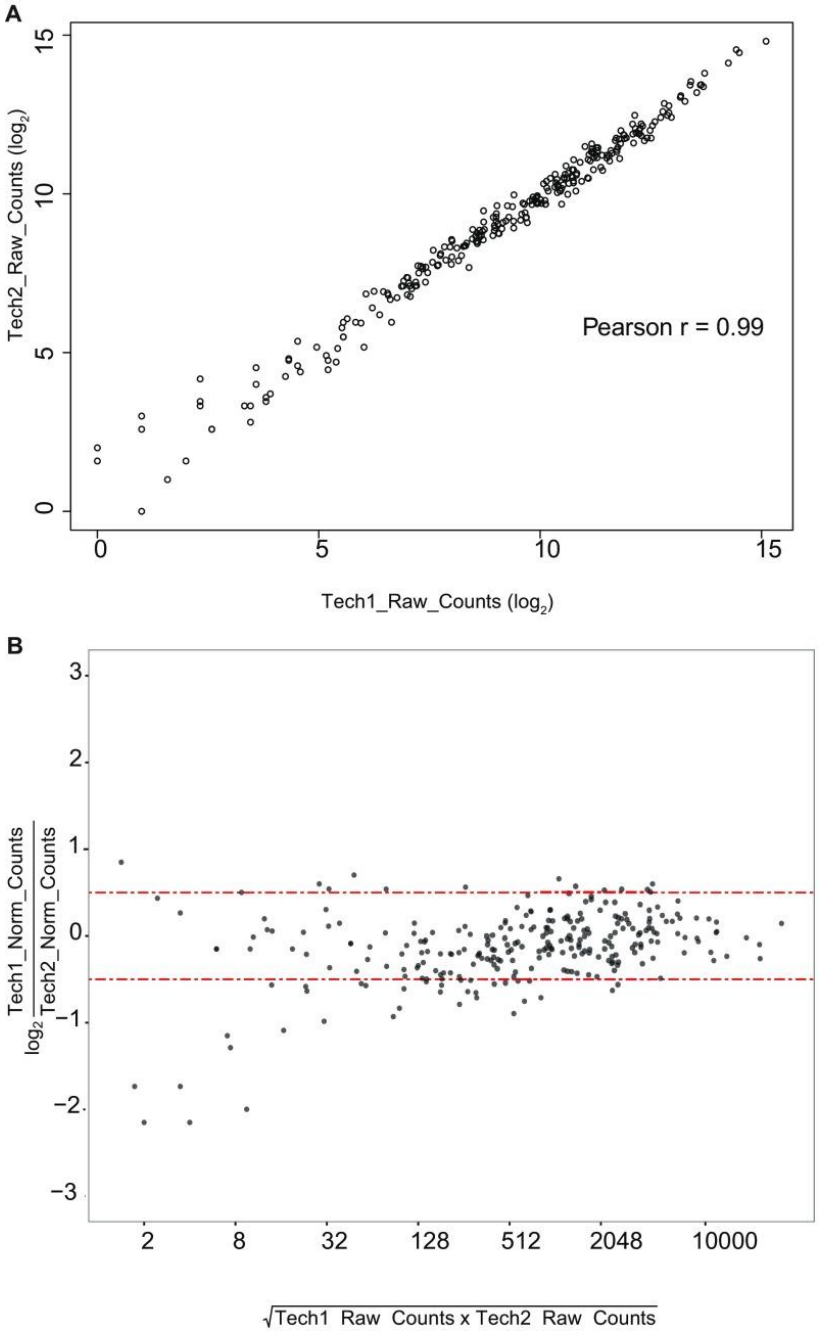


Figure 2.9 Lariat sequencing datasets are highly reproducible. (*A*) Raw counts for the 276 annotated introns used in the analysis in Figure 2.8*A* for two technical replicates of *prp8* N1869D are plotted as an XY scatter. (*B*) MA plot of the same data depicted in *A*. The *y* axis uses raw counts normalized to total raw counts from all annotated introns. Red lines mark changes ±0.5 (log_2) between the replicates.

For each event, the frequency of aberrant splice-site activation relative to the frequency of annotated splice-site activation was determined for the selected *prp8* toggle alleles. Figure 2.8A shows a false-colored representation of the behavior of each of these aberrant splicing events relative to *PRP8*. When considering the behavior of the different alleles across all of these splicing events, the behaviors of the pair of hyperaccurate/inefficient alleles highly correlated with one another (Pearson $r = 0.82$), and the behaviors of the pair of error-prone/efficient alleles highly correlated with one another (Pearson $r = 0.87$). By contrast, the behavior of the hyperaccurate/inefficient pairs was poorly correlated with that of the error-prone/efficient pairs (Pearson $r = -0.11$), consistent with these classes of alleles having opposing impacts on these aberrant events. Moreover, when considering the behavior of the individual splicing events in the context of *prp8* toggle alleles, events largely matched the results seen with *ACT1-CUP1* reporters (Figure 2.2 and Figure 2.3) wherein the hyperaccurate/inefficient alleles showed lower levels of aberrant splice-site activation than the WT, while the error-prone/efficient alleles showed higher levels. As with the reporters, the level of aberrant splice-site activation varied with the different *prp8* toggle alleles.

To independently validate the frequencies of specific aberrant splice-site activation events determined from lariat sequencing, we used a PCR-based approach to quantify the different lariat species (Figure 2.10). As seen in Figure 2.8*B*, primers were designed such that they faced away from one another in the context of linear RNA but toward one another in the context of a lariat. The amplicons derived from these primer pairs differed in length depending upon the particular 5′SS used in the reaction, allowing the relative abundancies of the two isoforms to be determined by capillary electrophoresis on a Bioanalyzer. The results

of these experiments were largely consistent with lariat sequencing results (Figure 2.8 *C–E*). Decreased levels of aberrant splice-site activation were apparent in the hyperaccurate/inefficient alleles relative to WT, with the strongest phenotypes apparent in both assays for the V1860D mutant, consistent with our observations on *ACT1-CUP1* reporters. The effects of the error-prone/efficient alleles were more modest, showing subtle yet consistent increases in the use of aberrant splice sites within some of the transcripts and little change from WT within others.

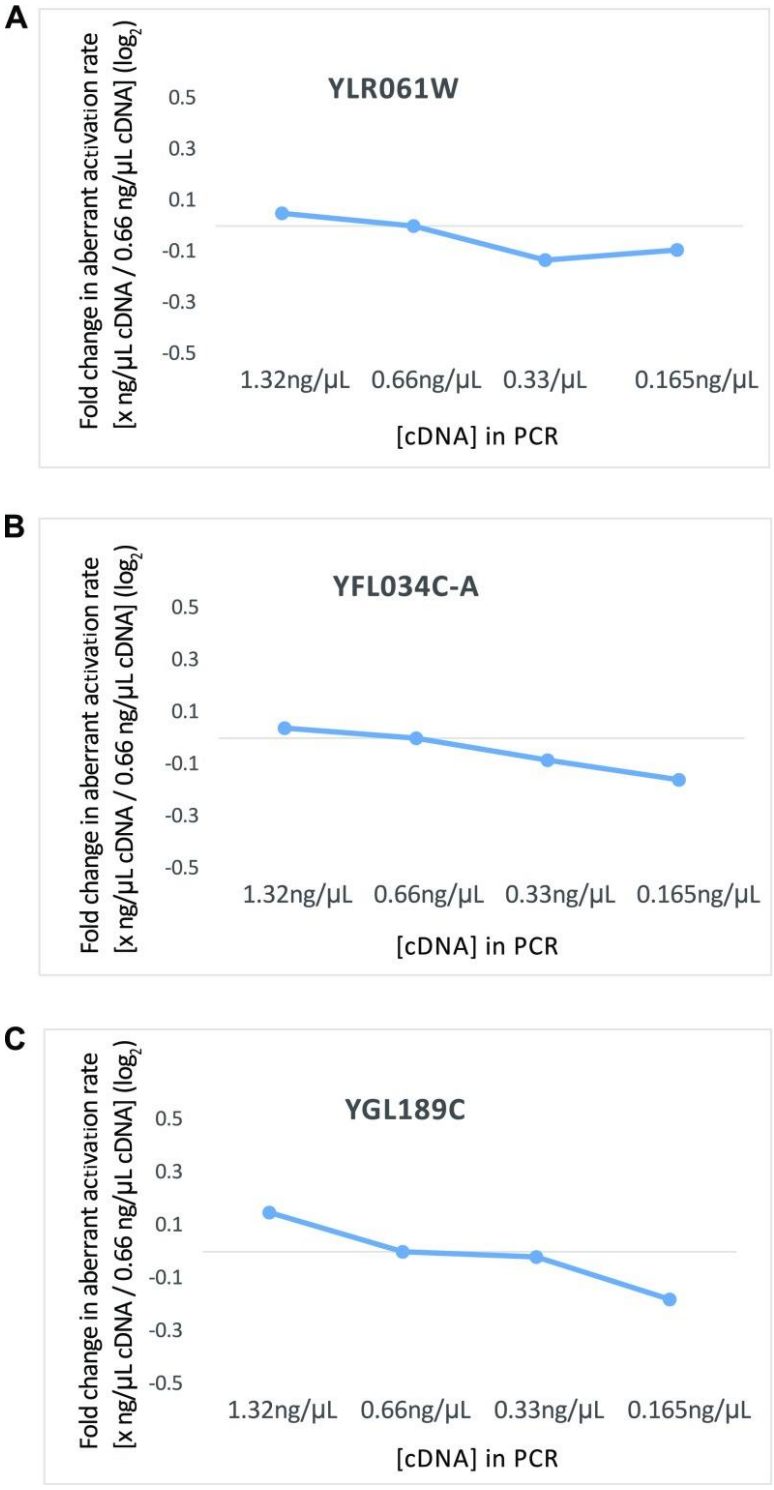


Figure 2.10 (*A*–*C*) Fold change in the measurement of aberrant activation rate with different concentrations of cDNA in the PCR over the concentration (0.66 ng/µL) that was used in the assay in Fig. 2.8 *C*–*E* are shown for the three introns investigated.

DISCUSSION

Here we present an analysis of the role of the RNaseH domain (RH) of Prp8 in spliceosomal activity. Specifically we reveal two classes of mutants in the RH extension of Prp8, transitional (inefficient/hyperaccurate) and catalytic (efficient/error-prone). Furthermore, because these transitional and catalytic *prp8* alleles stabilize mutually exclusive β-hairpin and loop conformations of the RH extension of Prp8 (Fig. 2.1 and Table 2.1) (19), our data directly link a conformational change in the spliceosome to genomewide changes in splicing fidelity.

The Toggle Model

We suggest that the RH extension of Prp8 is a structural toggle, oscillating between mutually exclusive conformations at distinct points in the splicing cycle to promote high-fidelity, high-efficiency splicing. Specifically, we propose that the RH extension adopts the transitional conformation before the first step, toggles to the catalytic for the first step, toggles back to the transitional between catalytic steps, and then again adopts the catalytic conformation for the second step—a "toggle model" (Fig. 2.11). That the RH extension adopts the catalytic form at both catalytic steps is supported by growth assays where yeast expressing catalytic *prp8* alleles promote the splicing of a variety of *ACT1-CUP1*reporters in vivo, regardless of which catalytic step is affected by the reporter (Figure 2.2 and Figure 2.3) (13, 16–19), our in vitro splicing assays (Fig. 2.5), and the structure of a B^{act} spliceosome (5). That RH adopts a transitional form before the first and again between the first and second steps is supported by structures of C and C* spliceosomes (7, 8, 10, 11) and by genetic interactions between *prp8* toggle alleles and the DEAH-box ATPases Prp2 and Prp16, which

promote spliceosome conformational changes required for the first and second steps. Catalytic *prp8* alleles suppress growth defects in yeast expressing ATPase-deficient Prp2, whereas transitional *prp8* alleles exacerbate them (2.4). These same catalytic alleles also suppress the growth defect of yeast expressing ATPase-deficient Prp16 (17, 18).

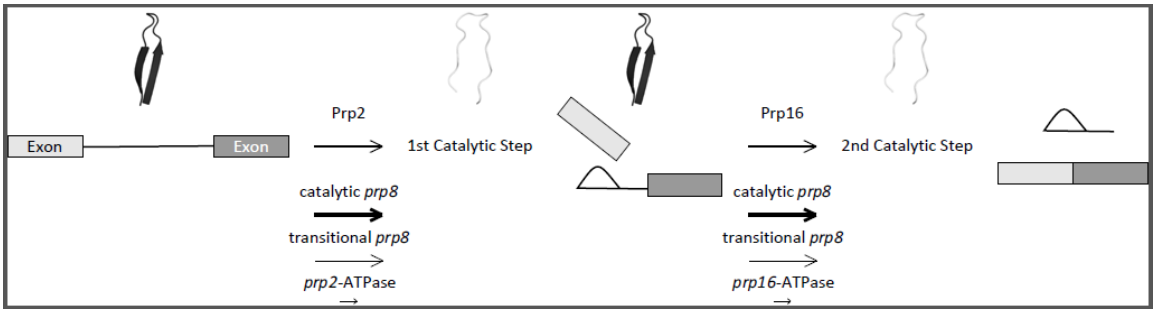


Figure 2.11 The toggle model. The Prp8 RH extension toggles between catalytic (error-prone/efficient) and transitional (hyperaccurate/inefficient) forms during the splicing cycle. Prp2 ATPase activity promotes the first catalytic step; Prp16 ATPase activity promotes the second. Catalytic *prp8* toggle alleles bias the spliceosome toward the catalytic conformation at both the first and second steps, whereas transitional mutants bias against conversion to the catalytic conformation.

The biological relevance of conformational toggling is well understood for the ribosome, another large, high-fidelity RNP machine that shares numerous similarities with the spliceosome (2). The ribosome repeatedly moves between "open" and "closed" conformations. Proofreading, which allows for the rejection of a near-cognate tRNA, takes place in the open state, whereas peptide bond catalysis occurs in the closed (35). Translational fidelity and ribosome conformation have been directly linked through studies

of ribosome mutants known as ribosomal ambiguity (*ram*) and restrictive. *Ram* mutants favor the closed, catalytic form of the ribosome and destabilize the proofreading conformation, even in the presence of near-cognate tRNA, resulting in coding errors and more rapid translation. Restrictive mutants have the opposite effect: They favor the open, proofreading conformation and are hyperaccurate (35).

The first-step/second-step model was proposed in 2004 and was the first spliceosomal model to invoke conformational toggling by the ribosome (17). By this model the spliceosome alternates between unique first- and second-step conformations (13, 17–19, 36) wherein first- and second-step alleles of *prp8* bias the conformation of the spliceosome toward first- or second-step conformations, respectively. Primer extensions performed on *ACT1-CUP1* reporter RNA isolated from strains carrying first- and second-step alleles were key to the development of this model. However, for the subset of *prp8* alleles that map to the RH extension, we note that these data are also consistent with the toggle model. Rather than a preference to dwell in a second-step conformation, decreased lariat intermediate and increased mRNA levels exhibited by second-step alleles can also indicate more rapid progression through both steps of splicing, as we propose. Similarly, instead of a bias toward a first-step conformation, decreased pre-mRNA and increased lariat intermediate levels observed in first-step alleles are consistent with a general bias against catalysis; the limited number of spliceosomes that succeed at the first step are unable to complete the second.

The RH domain of Prp8 is not the only spliceosomal component that toggles. Stem II of the U2 snRNA can fold into two mutually exclusive structures: Stem IIa and Stem IIc (37, 38). Alleles that bias toward the Stem IIa conformation have fidelity phenotypes similar

to *prp8* transitional alleles. These alleles also have genetic and proofreading phenotypes comparable to *prp16* ATPase-deficient alleles. In contrast, the U2 Stem IIc conformation has been shown to be necessary for both catalytic steps. Alleles that promote the U2 Stem IIc conformation have phenotypes similar to the *prp8* catalytic alleles, as do mutations within the myb-domain of Cef1 (36–38). We speculate that these individual toggles might be linked; the spliceosome as a whole may toggle between catalytic and transitional conformations.

CONCLUSION

Although the ribosome and spliceosome are separated by billions of years of evolution, they exhibit many similarities including our demonstration here of a direct coupling of catalytic efficiency with fidelity. Whereas work over the past decade has shed light on the mechanisms by which structural changes in the ribosome enable discrimination of cognate from near-cognate tRNAs, we have only begun to investigate these mechanisms in the spliceosome, owing at least in part to the difficulty in determining whether a splice site should be considered as "cognate" or "near-cognate." In fact, in the current work probing budding yeast, where splice sites have evolved to conform to a precise consensus sequence, the global locations of aberrant splice-site activation are marked by splice-site sequences that in many instances look far more cognate than do many annotated mammalian splice sites (Fig. 2.8 *C–E*). Nevertheless, there are many more locations across the genome where strong potential splice sites exist yet for which we detect no activation. Understanding why the spliceosome activates some of these sites but not others will be key to understanding the relationship between catalytic efficiency and fidelity on the spliceosome. Likewise, it will be crucial to identify structural conformations of the spliceosome that can trigger the

interconversion between proofreading and catalytic states, which could contribute to alternative splice-site selection in higher organisms.

MATERIALS AND METHODS

Strains and Plasmids.

See Table 2.3 for strain genotypes. In vitro splicing extracts were created from strain yTB72 (*Mat a prp8Δ::LYS2 his3 leu2 ura3 lys2 yCP50-PRP8*) (25) with transitional and catalytic allele variants shuffled in. Strain yAMP24 (*mat α cup1Δ::ura3-52 prp8Δ::LYS2 ade2 his3 leu2 ura3 lys2 yCP50-PRP8*) (40) was used in the *ACT1-CUP1* splicing reporter assays. Transitional and catalytic allelic variants of *prp8* were created by templated mutagenesis of a pRS313-*PRP8* plasmid using the KOD polymerase system (Novagen) and introduced to yTB72 or yAMP24 by 5FOA shuffling (39). The *ACT1-CUP1* reporter plasmids have been described previously (25). The strain yTT516 (*Mat a cup1Δ::ura3-52 prp2Δ::TRP yCP51-PRP2 prp8Δ::LYS yCP50-PRP8 ade leu his ura*) was created by mating stain yTB72 with strain yTY1 (41) and shuffling in plasmids pJPS1910, pJPS1919, and pJPS2501 (27), containing *PRP2* or *prp2-Q548E* encoded on a pRS415 vector (gifts from J. Staley, University of Chicago, Chicago). Strain yTT97 (*Mat a prp8Δ::LYS2 his3 leu2 ura3 lys2 dbr1::Kan*) was created from yTB72 by using homologous recombination to replace the coding sequence of *DBR1* with the Kan marker and was used for lariat sequencing analyses. Prp8 plasmids are available through Addgene (www.addgene.org).

Table 2.3 Genotype of strains used in this study

Strain	Genotype
yTB72	***mat a prp8Δ::LYS2 his3 leu2 ura3 lys2 yCP50-PRP8***
yAMP24	***mat α cup1Δ::ura3-52 prp8Δ::LYS2 ade2 his3 leu2 ura3 lys2 yCP50-PRP8***
yTT516	***mat a cup1Δ::ura3-52 prp2Δ::TRP yCP51-PRP2 prp8Δ::LYS yCP50-PRP8 ade leu his ura***
yTT97	***mat a prp8Δ::LYS2 his3 leu2 ura3 lys2 dbr1::Kan***

ACT1-CUP1 Reporter Assays.

Strains derivative of yAMP24 containing a *PRP8* allele with an *ACT1-CUP1* reporter plasmid were grown overnight to saturation, diluted back to OD_{600} of 0.1, then grown to OD_{600} of 0.5–0.8. Cultures were then diluted back to an OD_{600} of 0.1, serially diluted from there in fourfold dilutions, and spotted onto –Leu plates containing varying concentrations of $CuSO_4$ (0, 0.013, 0.025, 0.05, 0.075, 0.1, 0.18, 0.25, 0.3, 0.4, 0.5, 0.75, 1.0, and 1.5 mM). Plates were incubated at 30 °C for 3–4 d. The maximum concentration of copper a specific strain could grow on was assessed for each *prp8*mutant in combination with each *ACT1-CUP1* reporter and given a score calculated using the formula $\log_2([Cu^{2+}{}_{max}$ *prp8*-catalytic or transitional]/[$Cu^{2+}{}_{max}$ matched *PRP8-WT*]) to facilitate comparisons between strains and normalize between replicates (40). To capture minor differences between strains, if two strains of the same set (grown and spotted the same day onto the same batch of plates) were both able to tolerate the same maximum concentration of copper but one was growing clearly better or worse than the other at that concentration we adjusted the "tolerated copper" value

up or down (depending on growth relative to multiple other strains on the plate) by 10% before calculating growth relative to WT.

In Vitro Splicing Assays.

Yeast were grown in rich media at 30 °C until they reached an OD_{600} of ~1.0. Under these conditions, all mutants grew at a rate similar to *PRP8*. Extracts and pre-mRNA templates were made as described previously (25). In vitro splicing reactions consisted of a 40% vol/vol splicing extract combined with pre-mRNA template and 3% PEG 8000, 60 mM potassium phosphate, pH 7.0, 2.5 mM $MgCl_2$, 2 mM ATP, and 10 U RNasin (Promega). Reactions were performed as indicated (10 min at 37 °C for first-step analysis, 20 min at 30 °C for BrC template analysis) before being quenched with 2.5% SDS and 1 mM EDTA. Reactions were incubated with 10 U of proteinase K (Invitrogen) at 65 °C for 10 min and loaded [1:1 in formamide loading buffer (96% formamide, 20 mM EDTA, and bromophenol blue or no dye)] onto a 6% polyacrylamide (19:1), 7 M urea, Tris/borate/EDTA gel, visualized on a Typhoon imaging system (GE Healthcare Life Sciences), and quantified using ImageQuant (GE Healthcare Life Sciences). First-step product was quantified as $F_{lariat}/(F_{lariat} + F_{pre\text{-}mRNA})$, and the fraction spliced in the BrC mutant template assays was quantified as condition $(F_{lariat} + F_{lariat\text{-}intermediate} + F_{mRNA})/(F_{pre\text{-}mRNA} + F_{lariat} + F_{lariat\text{-}intermediate} + F_{mRNA})$. Values reported in Fig. 4 *C* and *D* represent the average and SEM for three distinct biological replicates performed using three separate preparations of splicing extract.

Lariat Sequencing: Library Preparation.

Total RNA isolation.

Cells were grown in yeast extract peptone dextrose (YPD) at 30 °C overnight to saturation in a 5-mL culture then back-diluted to an OD_{600} of 0.05 in a 25-mL YPD culture. When cells reached an OD_{600} between 0.4 and 0.8, 25 mL of cell culture was harvested by filtration then placed in a 15-mL Falcon tube and flash-frozen using liquid nitrogen and stored until RNA isolation. Total cellular RNA was isolated using hot phenol-chloroform extraction as follows. Into the 15-mL Falcon tube containing filter-collected cells was added 2 mL acid phenol (pH <5.5), followed by 2 mL AES buffer [50 mM sodium acetate (pH 5.3), 10 mM EDTA, and 1% SDS], and the tube was vortexed for 10 s then incubated at 65 °C for 7 min, vortexing for at least 3 s at 1-min intervals. The tube was then incubated on ice for 5 min, then the entire mixture was transferred to a 15-mL PhaseLock Heavy Gel tube (5PRIME) and centrifuged at 4 °C at 5,250 × *g* for 5 min. Two milliliters of phenol:chloroform:isoamyl alcohol (IAA) (25:24:1) was then added to the supernatant and mixed by shaking the tube up and down vigorously for 5 s. The tube was then centrifuged again in the same manner, and 2 mL chloroform was added to the supernatant, mixing as before. The tube was centrifuged as before, and this time the supernatant was transferred to a new 15-mL Falcon tube, to which 2.5 mL isopropanol and 200 μL 3 M sodium acetate (pH 5.3) were added. The resulting solution was mixed by vortexing and incubated at −20 °C for at least an hour. It was then centrifuged for 30 min at 4 °C at 5,250 × *g*, then supernatant was removed by decanting and the pellet was transferred to a 1.7-mL centrifuge tube. Then, the pellet was washed with 1.5 mL 70% ethanol twice by centrifuging for 5 min at 4 °C at 18,000 × *g*. The supernatant was

decanted, and the sample was dried using vacuum centrifugation at room temperature. The RNA pellet was then resuspended in RNase-free water.

DNase treatment.

Ten micrograms of total RNA isolated by the above method was then subjected to DNase treatment in a 20-µL reaction with a final composition of 40 mM Tris·HCl (pH 8.0), 10 mM $MgSO_4$, 1 mM $CaCl_2$, and 2 U of RQ1 RNase-free DNase (Promega). The reaction was incubated at room temperature for 15 min. The reaction was then cleaned up by phenol-chloroform extraction and ethanol precipitation. All subsequent steps in the protocol used Low Retention tubes (Fisherbrand) to minimize sample loss.

rRNA depletion.

DNased RNA (4.5 µg) was used as input for rRNA depletion using Illumina's Ribo-Zero Gold rRNA Removal Kit (Yeast) following the manufacturer's instructions. At the end of the protocol, 85 µL of supernatant was obtained, which was purified by ethanol precipitation. To the 85 µL supernatant, 15 µL of water, 10 µL of 3 M sodium acetate (pH 5.3), 250 µL of ice-cold 200-proof pure ethanol, and 2 µL of glycogen (20 g/L) were added and the solution was incubated at −20 °C for at least an hour and precipitated. The RNA pellet was resuspended in water.

RNaseR treatment.

RNaseR treatment was set up with rRNA-depleted RNA obtained from the previous step in a 50-µL reaction with a final composition of 20 mM Tris·HCl (pH 8.0), 100 mM KCl, 0.1 mM $MgCl_2$, and 1 U of RNaseR (Epicentre). The reaction was incubated at 37 °C for 10 min.

The reaction was then cleaned up by phenol-chloroform extraction and ethanol precipitation. The RNA pellet was then resuspended in water.

cDNA synthesis.

cDNA synthesis was performed with RNaseR-treated RNA using SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen) with modifications to the first-strand cDNA synthesis step. RNA and dN_9 primers were mixed in a 10 μL volume and incubated at 65 °C for 10 min followed by cooling on ice for 1 min and on benchtop for 5 min. To this sample, 4 μL of 5× First-Strand Reaction Buffer without $MgCl_2$ [250 mM Tris·HCl (pH 8.3) and 375 mM KCl] was added. Then, 2 μL 0.1 M DTT, 1 μL 10 mM dNTP, and 2 μL 30 mM $MnCl_2$ were added and the reaction was equilibrated at 45 °C for 2 min. Then, 1 μL of SuperScript II RT was added and the reaction was incubated at 45 °C for 1 h. As seen in Fig. S3, addition of $MnCl_2$ to a standard cDNA synthesis reaction with 3 mM $MgCl_2$ significantly increases the read-through frequency of the 2′,5′ linkage. Similar read-through efficiency was achieved using 3 mM $MnCl_2$ and no $MgCl_2$. The first-strand reaction was cleaned up with ethanol precipitation to remove $MnCl_2$ from the reaction. The pellet was then resuspended in 105 μL water, 4 μL 5× First-Strand Reaction Buffer with $MgCl_2$ [250 mM Tris·HCl (pH 8.3), 375 mM KCl, 15 mM $MgCl_2$], 2 μL 0.1 M DTT, 4 μL 10 mM dNTP, 30 μL 5× Second-Strand Reaction Buffer [100 mM Tris·HCl (pH 6.9), 450 mM KCl, 23 mM $MgCl_2$, 0.75 mM β-NAD+, and 50 mM $(NH_4)_2SO_4$], 4 μL *Escherichia coli* DNA Polymerase I (10 U/μL), and 1 μL *E. coli* RNase H (2 U/μL). *E. coli* DNA Ligase was not added. This reaction was incubated at 16 °C for 2 h. Following this, 2 μL of T4 DNA Polymerase (5 U/μL) was added and the reaction was further incubated at 16 °C for 5 min. The reaction was stopped by

addition of 10 μL of 0.5 M EDTA and then ethanol-precipitated. The double-stranded cDNA pellet was then resuspended in water and purified on a 10% native acrylamide gel and sized above 30 bp. The cDNA was then extracted from the gel by soaking the crushed gel pieces in four to five volumes of 0.3 M sodium acetate (pH 5.3) and rotating overnight. The eluate was then precipitated with 2.5 volumes of ethanol and 4 μL gycogen followed by two washes with 70% ethanol. The cDNA pellet was then resuspended in 85 μL water.

End repair.

To 85 μL of double-stranded cDNA from the previous step, 10 μL of 10× NEBNext End Repair Reaction Buffer (500 mM Tris·HCl, pH 7.5, 100 mM $MgCl_2$, 100 mM DTT, 10 mM ATP, and 4 mM dNTPs) and 5 μL of NEBNext End Repair Enzyme Mix (NEB End Repair Module) were added and the reaction was incubated at room temperature for 30 min. Cleanup was then done with phenol-chloroform extraction and ethanol precipitation. The RNA pellet was then resuspended in water. Further, the end-repaired DNA was cleaned up with GE Illustra MicroSpin G-25 columns per the manufacturer's instructions. It was then eluted in 50 μL water.

A-tailing.

To 42 μL of end-repaired DNA, 5 μL of 10× NEBNext dA-Tailing Reaction Buffer (100 mM Tris·HCl, pH 7.9, 100 mM $MgCl_2$, 0.5 M NaCl, 10 mM DTT, and 2 mM dATP) and 3 μL of Klenow Fragment (3′→ 5′ exo^-) (NEB dA-tailing module) were added and the reaction was incubated at 37 °C for 30 min. Following this, the reaction was cleaned up with phenol-

chloroform extraction and ethanol precipitation. Following precipitation, the pellet was resuspended in 16.5 μL water.

Adapter ligation.

Ligation was done with Enzymatics Rapid Ligation kit. To 16.5 μL of A-tailed DNA, 18.5 μL of 2× Rapid Ligation buffer [132 mM Tris·HCl (pH 7.6), 20 mM $MgCl_2$, 2 mM DTT, 2 mM ATP, and 15% PEG 6000], 1 μL of T4 DNA Ligase (Rapid) (L6030-HC-L), and 1 μL of Illumina Tru-seq barcoded adapter were added. This reaction was incubated at room temperature for 15 min.

Ligation sizing.

Thirty-seven microliters of 2× denaturing gel loading buffer (95% formamide and 25 mM EDTA) was added to the ligation reaction and this was run on 6% denaturing PAGE. The sample was sized between 150 nt and 330 nt. Ligated DNA was then extracted from the sized gel by soaking the crushed gel pieces in four to five volumes of 0.3 M sodium acetate (pH 5.3) and rotating overnight. The eluate was then precipitated with 2.5 volumes of ethanol and 4 μL glycogen followed by two washes with 70% ethanol and vacuum drying. The pellet was resuspended in 40 μL water.

PCR amplification.

Reactions were performed in 20 μL volume, with 10 μL DNA sample precipitated from gel, 1× Phusion HF buffer (NEB), dNTPs (0.25 mM each), 0.25 μM Illumina P5 PCR primer, 0.25 μM Illumina P7 PCR primer, and Phusion polymerase. The PCR protocol was an initial denaturation step of 98 °C for 30 s, followed by 8–10 cycles of 98 °C for 10 s, 65

°C for 30 s, and 72 °C for 30 s. Final extension was at 72 °C for 5 min. The PCR product was purified on a 6% native acrylamide gel, recovering the material between 150 and 330 bp. The purified products were sequenced on an Illumina NextSeq platform.

Cleanup by phenol-chloroform extraction and ethanol precipitation.

For phenol-chloroform extraction, the sample was made up to 100 μL or 200 μL with water and then an equal volume of phenol:chloroform:IAA (25:24:1) was added, mixed by shaking and centrifuged in a 2-mL PhaseLock Heavy Gel tube (5PRIME) at 18,000 × *g* for 2 min. Another volume of phenol:chloroform:IAA was again added to the supernatant and centrifuged again and then one volume of chloroform was added to the supernatant. The tube was centrifuged again and the supernatant was transferred to a 1.5-mL tube. To the supernatant, 1/10 volume of 3 M sodium acetate (pH 5.3), 2.5 volumes of ice-cold 200-proof pure ethanol, and 2 μL of glycogen were added and the solution was incubated at −20 °C for at least an hour. This was followed by centrifuging the tube at 18,000 × *g* for 30 min at 4 °C and removal of the supernatant. This was followed by two washes, each with 1 mL of 70% ethanol and centrifuging at 18,000 × *g* for 5 min. After the second wash, the supernatant was decanted, and the sample was dried using vacuum centrifugation at room temperature.

Libraries.

Lariat sequencing libraries were prepared for one biological replicate for each strain discussed in Figure 2.8. Additionally, to assess the technical reproducibility of the method, two technical replicates were processed separately from the step of rRNA depletion using a single sample of DNased RNA isolated from *prp8* N1869D. The method was highly

technically reproducible (Pearson r = 0.99) (Figure 2.9). The alignment statistics for all libraries are listed in Dataset S2 (http://www.pnas.org/content/114/18/4739/tab-figures-data).

Lariat Sequencing: Data Processing.

Lariat sequencing data can be accessed from the NCBI GEO database (accession no. GSE96891).

Lariat sequencing genome alignment.

Illumina sequencing reads were trimmed of any 3′ adapter using Trimmomatic with the following parameters: 3:30:10 and MINLEN:18. Trimmed reads were aligned to the *S. cerevisiae* genome (Ensembl R64-1-1) using Bowtie2. Parameters of alignment were "score-min L,-0.4,-0.4 –very-sensitive." End-to-end alignments were done for the initial genome alignment. Paired end alignments on lariat reads were done using "–ff -I 0 -X 5000 –no-mixed –no-discordant."

Branch-spanning lariat read alignment.

For identification of lariat reads spanning the lariat branch, each read that failed to align to the genome in end-to-end fashion was considered a candidate lariat branch read. These reads were split into mate paired reads at every GT dinucleotide in the forward and reverse complement strands. The subset of mate pairs where each read in the mate pair is at least 10 nucleotides were candidates for alignment. These reads were then assessed for alignment using Bowtie2 in paired-end mode with the fragment containing the GT as mate 1 and the other as mate 2, using the previously noted parameters. Because a single read split in this

fashion can yield several possible alignments, the possible GT split alignments were collapsed into a single best available paired-end alignment defined as the alignment that minimizes the number of mismatches.

Identification of annotated introns.

5′SS, Br locations, and strandedness were obtained from each lariat read using the position of GT on the lariat read alignment. Because mapped reads in the SAM/BAM format are represented on the forward strand of the reference genome, if the lariat read showed a GT dinucleotide in the beginning of the left part of the alignment the position of G in the genome was marked as the 5′SS with sense directionality and the last position of the right part was marked as the Br location. If the lariat read showed an AC dinucleotide in the last two positions of the right part of the alignment, the position of C in the genome was marked as the 5′SS with antisense directionality and the first position of the left part was marked as the Br location. Lariat reads with both a GT in the beginning of the left part and AC at the end of the right part were discarded as ambiguous reads (<0.2% of lariat reads). The 5′SS locations were then matched to the annotated 5′SS positions (Ensembl R64-1-1) to identify annotated introns. The list of identified annotated introns (287) along with their Br positions and counts for the *PRP8* WT dataset is shown in Dataset S1, A. For the analysis in Fig. 5*A*, only those annotated intronic reads with Br position mapping to annotated Br locations (based on Br-3′SS distance information from the Ares Intron Database 4.1, intron.ucsc.edu/yeast4.1/) were used (Dataset S1, B) (276 introns in the *PRP8* WT dataset). Because reverse transcriptase frequently introduces deletions and mutations when creating the cDNA product that crosses the 5′SS to the Br, to account for errors introduced during reverse transcription the identified

Br was allowed to be within ±2 nt of the annotated location to be characterized as a read containing annotated Br. Counts for each annotated intron were then the sum of all reads for that 5′SS with a Br position within ±2 nt of the annotated Br site.

Identification of aberrant 5′ splice sites.

To identify aberrant 5′SSs, only those lariat reads with Br position mapping to ±2 nt of the annotated Br locations were considered for analysis to reduce false-positive rate and restrict the analysis to aberrant 5′SS events in annotated introns. For these reads, if the 5′SS location was found to be more than 5 nt and less than 1,000 nt from the annotated 5′SS location, it was considered an aberrant 5′SS for that intron. The list of identified aberrant 5′SSs and their counts for the *PRP8* WT dataset are in Dataset S1, C.

Quantification of aberrant activation rate.

For each annotated intron, reads mapping to all of its aberrant 5′SSs were collected and their counts summed. Introns with collective aberrant counts less than 30 were excluded from further analysis because of the noise associated with low-count events. For the remaining events, the proportion of aberrant site use was calculated as

Aberrant activation rate= (Collective aberrant count for IntronXAnnotated counts for IntronX +Collective aberrant count for IntronX).

Validation of Aberrant Splice Sites by RT-PCR.

cDNA was generated from 1 μg of DNase-treated RNA using the methodology described above. PCR was done using primers (Table 2.4) designed to capture the annotated and most abundant aberrant splicing isoforms as described in Fig. 5*B*, with 3 min of initial

denaturation at 95 °C followed by 24–30 cycles of 95 °C for 10 s, 55 °C for 20 s, and 72 °C for 30 s in a 15-μL reaction containing 10 ng cDNA, 250 nM primers, 10 mM Tris·HCl (pH 8.5), 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM dNTPs (each), 0.25× SyBr Green, 5% vol DMSO, and Taq. Then, 1 μL of the reaction was run on an Agilent Bioanalyzer DNA 1000 assay to resolve the products and they were quantified with Agilent 2100 Bioanalyzer software. For YGL189C, because the aberrant 5′SS isoform was low in abundance relative to the annotated isoform, 10 such PCR reactions were run, pooled, purified, and concentrated into 10-μL volume using DNA Clean & Concentrator-5 kit from Zymo. One microliter of the purified product was then run on Agilent Bioanalyzer DNA 1000 assay. The area under the curve (AUC) for the bands corresponding to the annotated and most abundant aberrant isoform was used for the calculation of aberrant activation rate:

Aberrant activation rate= (AUC for the aberrant isoform AUC for the annotated isoform+AUC for the aberrant isoform).

To investigate whether the changes observed in aberrant activation rate in the mutants in Fig. 2.8 *C–E* are reflective of actual biological changes and not a result of technical artifacts, the aberrant activation rate for a given intron was calculated across a 16-fold dilution series of cDNA template in the PCR (Figure 2.10). Whereas YLR061W and YFL034C-A gave consistent results across a 16-fold cDNA dilution, YGL189C showed a 0.32-fold difference between the highest and lowest dilutions on a $\log_2$ scale. However, this change was considerably lower than the decreases in aberrant activation rate observed for V1860D and V1860N in Figure 2.8*E*.

Table 2.4 Sequences of RT-PCR Primer primers used

Primer	Sequence
YFL034C-A_FWD_LAR	GGCCAGACATTTTTTCCTCGTCC
YFL034C-A_RC_LAR	GTTCGCTGTGAAAGCGGGACTGTTC
YLR061W_FWD_LAR	GTAGGCAACTTTGTGGTTTCGGGA
YLR061W_RC_LAR	GCAACTAAACACATAGACGCTTC
YGL189C_FWD_LAR	CCTTCCCTTTTTGCCACGATC
YGL189C_RC_LAR	CAATATTCTTCAATAATAAGTAC

REFERENCES

1. Wahl MC, Will CL, Lührmann R. The spliceosome: Design principles of a dynamic RNP machine. Cell. 2009;136:701–718.

2. Semlow DR, Staley JP. Staying on message: Ensuring fidelity in pre-mRNA splicing. Trends Biochem Sci. 2012;37:263–273.

3. Nguyen THD, et al. Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution. Nature. 2016;530:298–302.

4. Wan R, et al. The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. Science. 2016;351:466–475.

5. Rauhut R, et al. Molecular architecture of the Saccharomyces cerevisiae activated spliceosome. Science. 2016;353:1399–1405.

6. Yan C, Wan R, Bai R, Huang G, Shi Y. Structure of a yeast catalytically activated spliceosome at 3.5 Å resolution. Science. 2016;353:904–11.

7. Galej WP, et al. Cryo-EM structure of the spliceosome immediately after branching. Nature. 2016;537:197–201.

8. Wan R, Yan C, Bai R, Huang G, Shi Y. Structure of a yeast catalytic step I spliceosome at 3.4 Å resolution. Science. 2016;353:895–904.

9. Yan C, et al. Structure of a yeast spliceosome at 3.6-angstrom resolution. Science. 2015;349:1182–1191.

10. Bertram K, et al. Cryo-EM structure of a human spliceosome activated for step 2 of splicing. Nature. 2017;542:318–323.

11. Fica SM, et al. Structure of a spliceosome remodelled for exon ligation. Nature. 2017;542:377–380.

12. Grainger RJ, Beggs JD. Prp8 protein: At the heart of the spliceosome. RNA. 2005;11:533–557.

13. Yang K, Zhang L, Xu T, Heroux A, Zhao R. Crystal structure of the beta-finger domain of Prp8 reveals analogy to ribosomal proteins. Proc Natl Acad Sci USA. 2008;105:13817–13822.

14. Kuhn AN, Brow DA. Suppressors of a cold-sensitive mutation in yeast U4 RNA define five domains in the splicing factor Prp8 that influence spliceosome activation. Genetics. 2000;155:1667–1682.

15. Kuhn AN, Reichl EM, Brow DA. Distinct domains of splicing factor Prp8 mediate different aspects of spliceosome activation. Proc Natl Acad Sci USA. 2002;99:9145–9149.

16. Collins CA, Guthrie C. Allele-specific genetic interactions between Prp8 and RNA active site residues suggest a function for Prp8 at the catalytic core of the spliceosome. Genes Dev. 1999;13:1970–1982.

17. Query CC, Konarska MM. Suppression of multiple substrate mutations by spliceosomal prp8 alleles suggests functional correlations with ribosomal ambiguity mutants. Mol Cell. 2004;14:343–354.

18. Liu L, Query CC, Konarska MM. Opposing classes of prp8 alleles modulate the transition between the catalytic steps of pre-mRNA splicing. Nat Struct Mol Biol. 2007;14:519–526.

19. Schellenberg MJ, et al. A conformational switch in PRP8 mediates metal ion coordination that promotes pre-mRNA exon ligation. Nat Struct Mol Biol. 2013;20:728–734.

20. Siatecka M, Reyes JL, Konarska MM. Functional interactions of Prp8 with both splice sites at the spliceosomal catalytic center. Genes Dev. 1999;13:1983–1993

21. Fica SM, et al. RNA catalyses nuclear pre-mRNA splicing. Nature. 2013;503:229–234.

22. Marcia M, Pyle AM. Visualizing group II intron catalysis through the stages of splicing. Cell. 2012;151:497–507.

23. Umen JG, Guthrie C. Mutagenesis of the yeast gene PRP8 reveals domains governing the specificity and fidelity of 3′ splice site selection. Genetics. 1996;143:723–739.

24. Lesser CF, Guthrie C. Mutational analysis of pre-mRNA splicing in Saccharomyces cerevisiae using a sensitive new reporter gene, CUP1. Genetics. 1993;133:851–863.

25. Mayerle M, Guthrie C. Prp8 retinitis pigmentosa mutants cause defects in the transition between the catalytic steps of splicing. RNA. 2016;22:798–809.

26. Ohrt T, et al. Prp2-mediated protein rearrangements at the catalytic core of the spliceosome as revealed by dcFCCS. RNA. 2012;18:1244–1256.

27. Wlodaver AM, Staley JP. The DExD/H-box ATPase Prp2p destabilizes and proofreads the catalytic RNA core of the spliceosome. RNA. 2014;20:282–294.

28. Anderson K, Moore MJ. Bimolecular exon ligation by the human spliceosome bypasses early 3′ splice site AG recognition and requires NTP hydrolysis. RNA. 2000;6:16–25.

29. Fouser LA, Friesen JD. Mutations in a yeast intron demonstrate the importance of specific conserved nucleotides for the two stages of nuclear mRNA splicing. Cell. 1986;45:81–93.

30. Taggart AJ, DeSimone AM, Shih JS, Filloux ME, Fairbrother WG. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. Nat Struct Mol Biol. 2012;19:719–721.

31. Awan AR, Manfredo A, Pleiss JA. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. Proc Natl Acad Sci USA. 2013;110:12762–12767.

32. Stepankiw N, Raghavan M, Fogarty EA, Grimson A, Pleiss JA. Widespread alternative and aberrant splicing revealed by lariat sequencing. Nucleic Acids Res. 2015;43:8488–8501.

33. Gould GM, et al. Identification of new branch points and unconventional introns in Saccharomyces cerevisiae. RNA. 2016;22:1522–1534.

34. Qin D, Huang L, Wlodaver A, Andrade J, Staley JP. Sequencing of lariat termini in S. cerevisiae reveals 5′ splice sites, branch points, and novel splicing events. RNA. 2016;22:237–253.

35. Ogle JM, Murphy FV, Tarry MJ, Ramakrishnan V. Selection of tRNA by the ribosome requires a transition from an open to a closed form. Cell. 2002;111:721–732.

36. Query CC, Konarska MM. CEF1/CDC5 alleles modulate transitions between catalytic conformations of the spliceosome. RNA. 2012;18:1001–1013.

37. Hilliker AK, Mefford MA, Staley JP. U2 toggles iteratively between the stem IIa and stem IIc conformations to promote pre-mRNA splicing. Genes Dev. 2007;21:821–834.

38. Perriman RJ, Ares M., Jr Rearrangement of competing U2 RNA helices within the spliceosome promotes multiple steps in splicing. Genes Dev. 2007;21:811–820.

39. Guthrie C, Fink GR, editors. Guide to Yeast Genetics and Molecular and Cell Biology. Elsevier; London: 2004.

40. Price AM, Görnemann J, Guthrie C, Brow DA. An unanticipated early function of DEAD-box ATPase Prp28 during commitment to splicing is modulated by U5 snRNP protein Prp8. RNA. 2013;20:40–60.

41. Edwalds-Gilbert G, et al. Dominant negative mutants of the yeast splicing factor Prp2 map to a putative cleft region in the helicase domain of DExD/H-box proteins. RNA. 2000;6:1106–

1119.

42. Collins CA, Guthrie C. Genetic interactions between the 5′ and 3′ splice site consensus sequences and U6 snRNA during the second catalytic step of pre-mRNA splicing. RNA. 2001;7:1845–1854.

CHAPTER 3

INVESTIGATING THE RELATIONSHIP BETWEEN PRP8 AND SPLICE SITE SELECTION: HAS THE RH EXTENSION WITHIN THE RNASEH DOMAIN OF PRP8 EVOLVED TO DICTATE STRINGENCY?

ABSTRACT

The nature of splice sites in spliceosomal introns has differentially evolved across eukaryotes despite the deep conservation in splicing mechanism and machinery. While most organisms have degenerate splice sites, similar to the Last Common Eukaryotic Ancestor (LECA), some organisms have evolved splice sites that conform to a rigid consensus. It is not completely clear which components of the splicing machinery co-evolved with these changes in splice site nature, nor even which of these changes occurred first. Working in the budding yeast, *S. cerevisiae*, an organism which has evolved to retain only a small number of introns, almost all with strong consensus splice site sequences, we recently demonstrated that mutations within the RNaseH domain of the core spliceosomal factor Prp8 resulted in anti-correlated changes in splicing fidelity and efficiency. In particular, mutations within the beta-finger extension of the RNaseH domain (RH extension) yield spliceosomes with increased or decreased capacities to activate non-canonical splice sites, concomitant with decreases or increases in splicing efficiency, respectively. To better understand the function of this region of Prp8, here I have investigated whether this region has differentially evolved in organisms with rigid splice sites in comparison to those with degenerate splice sites. I have investigated its evolution across a broad swath of organisms which display varying levels of degeneracy in the sequences of their splice sites. Remarkably, the RH extension residues are mostly invariant among widely diverged species across almost all domains of eukaryotic life, many of whom can be said to have degenerate splice site sequences. By contrast, in most organisms where the amino acid sequence of the RH extension has evolved, splice site sequences are seen to conform to a more rigid consensus, suggesting the fascinating

possibility that the RH extension is associated with the constriction of splice site sequences in these organisms.

INTRODUCTION

The protein-coding regions of almost all eukaryotic genes are interrupted by non-coding sequences called introns. Introns are removed from the transcribed pre-mRNA through the process of splicing, which is carried out by the macromolecular complex called the spliceosome. Two key cis elements play a primary role in intron recognition – the 5' splice site (5'SS) at the 5' boundary of the intron and the branch point (BP) near the 3' boundary of the intron. Recognition of these elements is largely through base pairing with the RNA components of the spliceosome. The strength of these elements, therefore, is largely determined by their sequence complementarity to the base pairing regions of the spliceosome (Sharp, 2005).

Splicing is a remarkably conserved process. Strong evidence exists in support of the model that eukaryotic spliceosomal introns evolved from group II introns, a class of self-splicing ribozymes found in almost all domains of life including bacteria. Even though the group II intron is an RNA complex and the spliceosome is a ribonucleoprotein (RNP) complex, the structure of the active site and the surrounding scaffold share remarkable similarities suggesting a deep conservation of the fundamental mechanism of splicing (Galej, Toor, Newman, & Nagai, 2018). Further, the core machinery of the spliceosome is highly conserved among eukaryotes (Anantharaman, Koonin, & Aravind, 2002; Collins & Penny, 2005).

Despite the deep conservation of the core splicing machinery, gene structures, including the strength of splice site sequences, have evolved differently in different organisms (Irimia, Penny, & Roy, 2007; Schwartz et al., 2008). Evolutionary studies of spliceosomal introns based on comparative genomics of extant eukaryotes as well as reconstruction of ancient

genomes strongly suggest that the Last Eukaryotic Common Ancestor (LECA) possessed an intron rich genome with weak signals and complex splicing machinery, as in present day humans, before subsequent divergence (Galej, Toor, Newman, & Nagai, 2018b; Irimia & Roy, 2014; Rogozin, Carmel, Csuros, & Koonin, 2012). From the ancestral state in the LECA with degenerate splice site sequences, most of the present day eukaryotes have retained the degenerate nature. However, a few organisms across different clades have evolved splice site sequences that conform to a rigid consensus (Schwartz et al., 2008). \

It is not completely clear what changes in the splicing machinery are associated with the rigidification process. Changes in auxiliary factors and/or the variable regions of the core splicing machinery that result in reduced efficiency of usage of degenerate sites could have coevolved with the rigidification of splice sites. Because splice site strength is primarily driven by the sequences of the splice sites, the efficiency of usage of degenerate splice sites can be viewed in terms of fidelity of splicing, with higher efficiency of degenerate site usage reflecting lower fidelity, where a wide variety of sequences are allowed as splice sites, and vice versa. The fidelity and efficiency of splicing have been shown to be influenced both at initial intron recognition as well as at the later catalytic steps of splicing (Koodathingal & Staley, 2013; Will & Lührmann, 2011). These are potential points where changes associated with rigidification could have happened.

While many studies have investigated the association of intron recognition machinery with splice site degeneracy, not much is known about the influence of the machinery involved at the catalytic stages of splicing. For instance, SR proteins, which have long been known to aid in recognition of weak splice sites, are not present in hemiascomycete fungi which possess rigid splice sites, suggesting that splice site rigidification in the

hemiascomycetes has been accompanied by the loss of SR proteins (Plass, Agirre, Reyes, Camara, & Eyras, 2008). Similarly, polypyrimidine tract, a pyrimidine-rich region found between the BP and the 3'SS, and factors that bind to it were found to be differentially evolved between organisms with degenerate and rigid splice sites, suggesting a connection between the polypyrimidine tract and the ability to use degenerate splice sites (Schwartz et al., 2008). While the above instances show the importance of intron recognition machinery in influencing usage of degenerate splice site sequences, most of this machinery is no longer present when catalysis happens. And, it is known that the substrates for catalysis are proofread even at later steps in the pathway. Therefore, it is intriguing whether the catalytic machinery also coevolved with splice site nature.

Prp8 is a core spliceosomal factor that acts as a scaffold for the active site of the spliceosme. It joins the spliceosome soon after assembly and stays until the two chemical steps of splicing take place (Fica & Nagai, 2017). Not surprisingly, considering its critical position in the spliceosome, it is the most highly conserved protein in the spliceosome, with 60% identity even between widely diverged plants and animals. Previous studies have shown that it is important for fidelity of splice site selection, with many suppressors of splice site mutations mapping to Prp8 (Grainger & Beggs, 2005). Many of these mutations are in Prp8's RNaseH (also called RNaseH-like) domain, and especially cluster around a 17-amino acid protrusion called the RH extension, that exists in two conformations - a β-hairpin and a freeform loop (Pena, Rozov, Fabrizio, Lührmann, & Wahl, 2008; Schellenberg et al., 2013; Yang, Zhang, Xu, Heroux, & Zhao, 2008). We showed recently that mutations that increase the relative stability of the hairpin conformation increase the fidelity and decrease the efficiency of splicing. In contrast, mutations that increase the relative stability of the loop

conformation increase the efficiency and decrease the fidelity of splicing (Mayerle et al., 2017).

Given that the relative stability of the hairpin/loop conformation of Prp8's RH extension is associated with altered fidelity/efficiency of the spliceosome, has the Prp8 RH extension evolved differently in organisms using rigid sites in comparison to those with degenerate sites and does this follow predictions from the conformational influence of RH extension on splicing nature? If the answer is yes, then organisms with rigid sites are expected to have a higher relative stability for the hairpin conformation of RH extension in comparison to organisms with degenerate sites.

Biochemical studies to directly assess the efficiency/fidelity of splicing and structural studies to characterize the structure of the RH extension in the different species would help in directly answering the question. However, before such elaborate studies are undertaken, as a first-step investigation, here I have compared the primary amino acid sequence and predicted secondary structure of Prp8 RH extension in 45 organisms which have rigid or degenerate splice sites and which are spread out across 4 super groups (domains) of eukaryotic life. Although the primary amino acid sequence by itself does not provide any indication of the structure of the RH extension, differential evolution of the primary amino acid sequence across rigid and degenerate groups is a strong indicator of a potential association between RH extension structure and splice site degeneracy that would help in designing more direct and targeted studies to test the association. Further, I have used secondary structure predictions to supplement the primary sequence analysis.

MATERIALS AND METHODS

Multiple sequence alignment of Prp8

Amino acid sequences of Prp8 for the organisms under study were obtained from UniProt. Multiple Sequence Alignment (MSA) was performed on these sequences with Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) using default parameters.

Splice site degeneracy calculation

Splice site degeneracy or rigidity is inferred from the information content of splice sites. The information content calculated by Irimia and co-workers (Irimia et al., 2007) was used for this analysis and it was derived as described in the paper: "Information content is a measure of the similarity among sequences, in this case 5′ splice site (5′ss) sequences within genomes, based on information theory. It reflects the certainty of finding a nucleotide in a given position of a consensus sequence, as a function of its entropy. The information content (IC) at a position, at which the letters A, C, G and T occur with frequencies pA, pC, pG and pT, respectively, is defined as IC = 2 + pA·log2pA + pC·log2pC + pG·log2pG + pT·log2pT, and ranges from 0 to +2 per nucleotide. Greater information contents will thus reflect greater similarity among sequences (i.e. species with stronger intron boundaries)." An IC of 4 was used as the boundary between the two groups, where less than 4 is termed as degenerate and greater than 4 is rigid (Figure 3.1)

.

Secondary structure prediction for Prp8 RH extension

Crystal structure of a part of human RNaseH domain containing the RH extension (PDB ID – 4jk7) was used as a model for structural analysis. There are two structures, where the

main difference is in the RH extension, which is either in hairpin or loop conformation. The structures were visualized with Chimera and all possible H-bonds were drawn using default parameters in Chimera (Tools -> Structure Analysis -> Find H-bonds). Then, using the Structure Editing tool, substitutions in RH extension were introduced using the Rotamers feature and high probability positions for the side group were investigated. At these positions, formation of any H-bonds or disruption of existing H-bonds was predicted with Chimera with default parameters. Any H-bonds formed or removed between the two strands of the hairpin were predicted to stabilize or destabilize the hairpin respectively.

RESULTS AND DISCUSSION

In order to investigate the evolution of Prp8's RH extension as a function of splice site nature, I characterized its primary amino acid sequence across a broad swath of organisms showing a range of splice site degeneracy. I leveraged work done by Irimia and co-workers, who sought to understand the relation between splice site degeneracy and intron number in the background of 49 eukaryotes with varying levels of splice site degeneracy (Figure 3.1) (Irimia et al., 2007) and used the organisms chosen in their study. These belonged to all four super groups/domains of eukaryotic life (Adl et al., 2012): Unikonta (metazoans, fungi, and amoebozoa), Archaeplastida (red algae, green algae and plants), SAR (alveolates, heterokonts and diatoms) and Excavata (members are unicellular organisms). Of the 49 organisms used by Irimia and co-workers, orthologs for Prp8 were found for 44 of the organisms. No orthologs were found for the following 5 species: *U. maydis, P. polycephalum, S. strix, R. americana, and M. jakobiformis*, and hence, these were left out of

the analysis. Within Unikonta, Archaeplastida and SAR, organisms containing either degenerate or rigid splice sites are present. However, there is only one excavate (*Trichomonas vaginalis)* with a Prp8 ortholog in this analysis and it has rigid splice sites. Although having a single organism in a domain does not provide information about the evolution of RH extension in that domain, as a stand-alone organism with rigid splice sites, it provides information on how the pool of organisms with rigid splice sites has evolved.

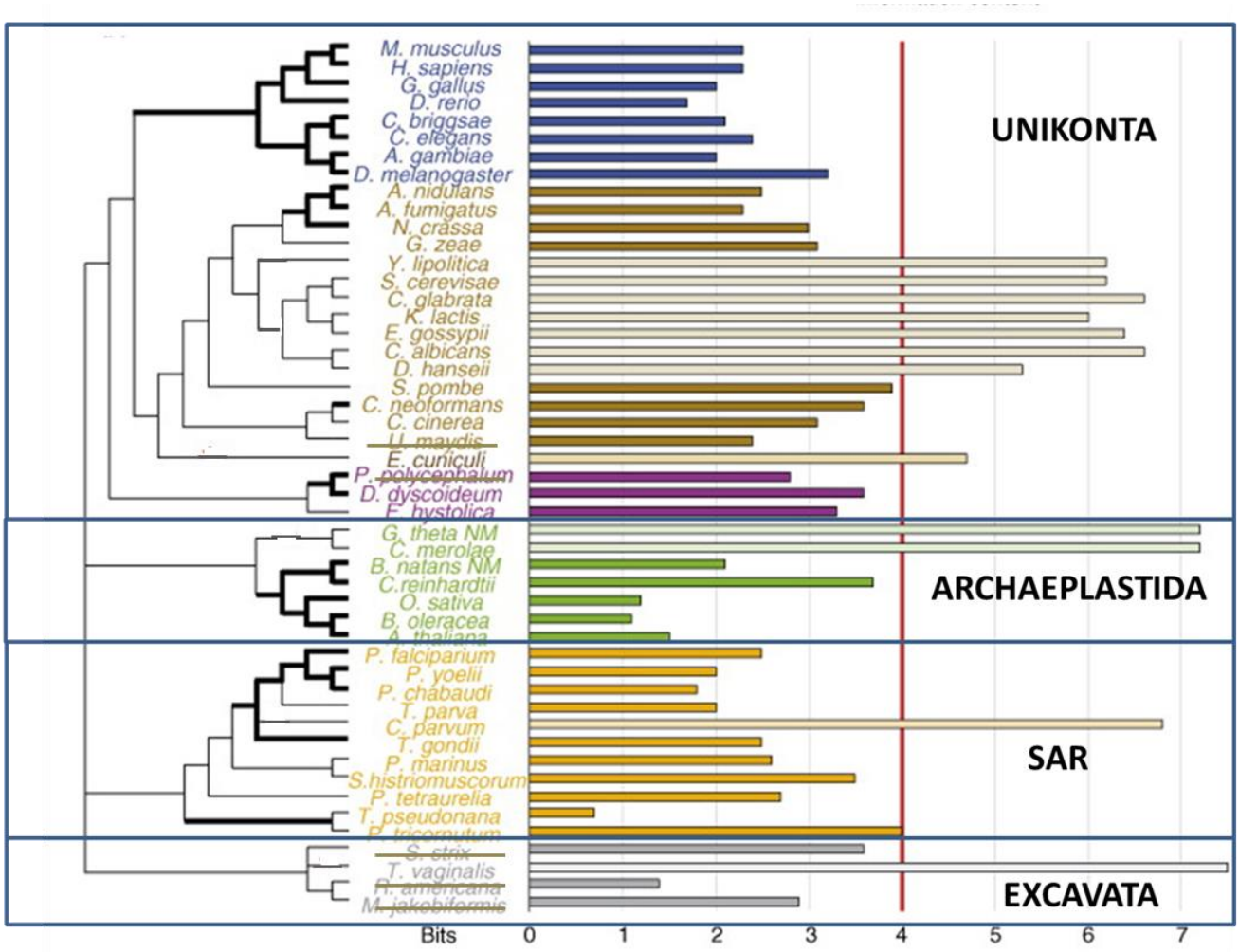


Figure 3.1 (modified from (Irimia et al., 2007)): On the left is the phylogenetic tree of the organisms used in the analysis. Plotted is the Information Content (IC) of the 5' splice sites in bits. The red line at 4 bits demarcates degenerate (<4 bits) and rigid nature (>4 bits). The names of organisms for which an annotated Prp8 ortholog or an ortholog retrievable from BLAST search was not available are struck-through. These were *U. maydis, P. polycephalum, S. strix, R. americana, and M. jakobiformis*.

Reconstruction of the ancestral Prp8 RH extension in LECA

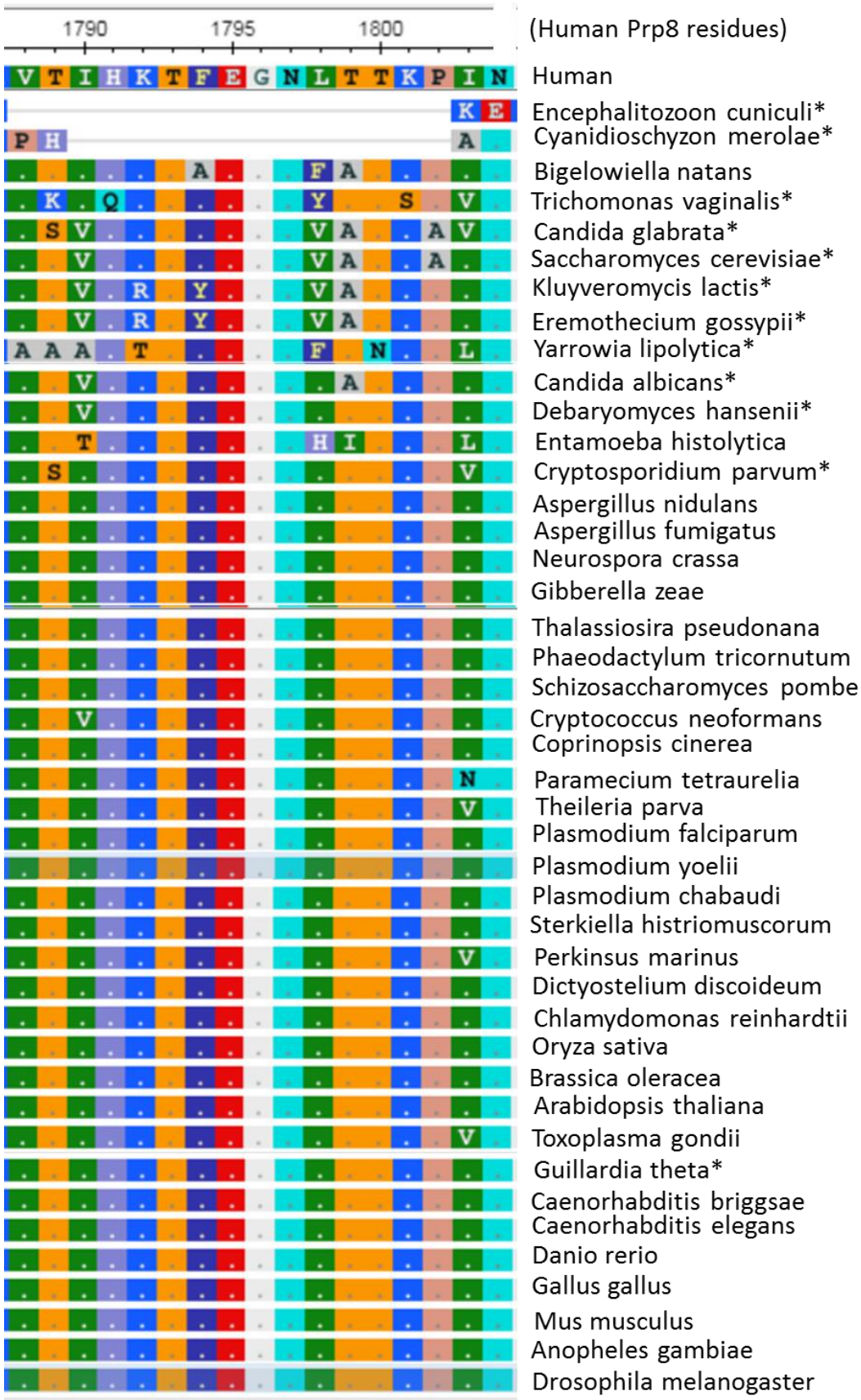
In order to understand how the RH extension has evolved across eukaryotes, the ancestral sequence in LECA was first reconstructed based on conservation of the region across eukaryotic groups. In the multiple sequence alignment of Prp8 orthologs (Figure 3.2), the RH sequence was identical in 27 species. The remaining 17 were different from one another and there was no single consensus. The 27 species with identical sequence included species from three super groups (excluding Excavata). Given that these super groups diverged after the LECA and the spliceosome was already present in the LECA, the presence of an invariant sequence across these groups suggests that the ancestral sequence in the LECA was most likely the same sequence.

Primary amino acid sequence of Prp8 RH extension is diverged in organisms with rigid splice sites in comparison to organisms with degenerate splice sites

According to a model where the RH extension in Prp8 in the LECA functioned to regulate the splicing of introns with degenerate splice sites, the prediction for its conservation in the extant eukaryotes under study is that those using degenerate sites are more likely to have retained this sequence. In contrast, those organisms that evolved to have rigid splice sites are more likely to have diverged from this sequence to acquire substitutions that presumably increased the relative stability of the hairpin conformation. Indeed, on comparison of rigid and degenerate groups in present day organisms, there is a strong correlation between divergence in RH extension sequence and rigidity of splice site usage.

In the three super groups (all except Excavata) in which species under study show splice site degeneracy like the LECA, the vast majority (25 of 32) are identical to the hypothesized ancestral sequence, while four display conservative substitutions. Only three have non-conservative substitutions. On the other hand, among the lineages which have rigidified across evolution, the majority (9 of 12) have at least one non-conservative substitution. This is true across three super families suggesting that several of the independent rigidification events investigated in this study are associated with divergence in RH extension sequence. These observations are largely consistent with the above model and suggest that there is differential evolution of the primary amino acid sequence of the RH extension of Prp8 across rigid and degenerate groups. This points to a potential differential at the structural level as well, as proposed by the hypothesis.

Figure 3.2: Multiple sequence alignment of the 17 amino acid Prp8 RH extension sequence in the 44 Prp8 orthologs. Human sequence is set as the master sequence with respect to which differences in other sequences are indicated. Identity to the human sequence is indicated as a dot. Any deviations from human residues are shown with the substituted amino acid single letter code. The amino acids are colored using a coloring scheme where each color corresponds to a group of amino acids with similar properties. So, a substitution by the same color is considered a conservative substitution and if otherwise, it is considered a non-conservative substitution. The hypothesized LECA sequence is identical to the human sequence. * - organisms categorized as using rigid splice sites are marked with this symbol.


1790
1795
1800
(Human Prp8 residues)
Human
Encephalitozoon cuniculi*
Cyanidioschyzon merolae*
Bigelowiella natans
Trichomonas vaginalis*
Candida glabrata*
Saccharomyces cerevisiae*
Kluyveromycis lactis*
Eremothecium gossypii*
Yarrowia lipolytica*
Candida albicans*
Debaryomyces hansenii*
Entamoeba histolytica
Cryptosporidium parvum*
Aspergillus nidulans
Aspergillus fumigatus
Neurospora crassa
Gibberella zeae
Thalassiosira pseudonana
Phaeodactylum tricornutum
Schizosaccharomyces pombe
Cryptococcus neoformans
Coprinopsis cinerea
Paramecium tetraurelia
Theileria parva
Plasmodium falciparum
Plasmodium yoelii
Plasmodium chabaudi
Sterkiella histriomuscorum
Perkinsus marinus
Dictyostelium discoideum
Chlamydomonas reinhardtii
Oryza sativa
Brassica oleracea
Arabidopsis thaliana
Toxoplasma gondii
Guillardia theta*
Caenorhabditis briggsae
Caenorhabditis elegans
Danio rerio
Gallus gallus
Mus musculus
Anopheles gambiae
Drosophila melanogaster

Secondary structure predictions to test differential evolution of RH extension structure in organisms with rigid and degenerate splice sites

To investigate if there is a differential in RH extension at the structural level between rigid and degenerate splicers, I next performed secondary structure predictions for the RH extension across organisms. I assessed how the rigid and degenerate pools have diverged structurally from LECA by making predictions for the effects of each of the observed non-conservative substitutions from the hypothesized LECA sequence in the diverged organisms. According to my hypothesis, the substitutions are predicted to increase the stability of the hairpin/loop conformation in the rigid/degenerate splicers respectively. The predictions were made on a human Prp8 RNaseH domain crystal structure (PDB ID – 4jk7) as described in the Methods section. The residue numbering mentioned in the following paragraphs is with respect to human Prp8 protein.

There are some caveats to this analysis: 1. the predictions were made for how each single substitution affects the structure. However, a composite of how all the different substitutions in a given organism is likely to change the structure is hard to predict and this should be kept in mind while interpreting the results. 2. This analysis assumes that the conservative substitutions do not contribute to any structural changes. However, it is possible that the conservative substitutions affect interactions with nearby partners and hence have an effect on the structure.

Therefore, the secondary structure predictions are only a preliminary attempt to understand the structural changes in the diverged organisms. If there are interesting differences between the rigid and degenerate splicers in this analysis, it warrants a more directed structural approach to study this.

Secondary structure predictions for non-conservative substitutions in organisms with degenerate splice sites suggest a weakening of the hairpin conformation of Prp8 RH extension

First, I investigated how the non-conservative substitutions affect the structure of the extension in organisms with degenerate splice sites. According to my hypothesis, the substitutions are predicted to increase the stability of the loop conformation in these organisms.

Of the organisms with degenerate sites, there are three that have diverged with non-conservative substitutions from the LECA sequence - the alga *B. natans,* the amoebozoan *E. histolytica,* and the alveolate *P. tetraurelia*. Based on secondary structure predictions, here are the predicted changes in the relative stability of hairpin/loop conformation in these three organisms.

E. histolytica – It has three non-conservative substitutions - T1799I, I1790T, L1798H.

T1799I – T1799 is involved in backbone as well as side chain H-bonding with H1791 which is across the hairpin. The change from T to I disrupts the sidechain bonding across the hairpin. This is highly likely to weaken the hairpin conformation.

I1790T - The I1790 residue is not involved in back-bone H-bonding in the hairpin structure. However, the residue lies close to T1789 which is involved in backbone and side chain H-bonding across the hairpin. The substitution from I to T at 1790 introduces a polar group in the structure which has the potential to be involved in external H-bonding. If this interaction is powerful, it has the potential to strain the adjacent hairpin-stabilizing H-bond involving T1789. This is likely to destabilize the hairpin conformation. However, it also has the

potential to form a bond across to the backbone. Therefore, it is hard to predict what the effect would be.

L1798H – L1798 is not involved in any H-bonding in the hairpin structure and the change to H also does not result in H-bonds within the structure. However, it is a drastic change from a non-polar group to a basic / aromatic side group which has the potential to be involved in inter-molecular interactions. It is difficult to predict whether this substitution will stabilize or destabilize the hairpin structure.

Overall, although the effects of L1798H and I1790T are difficult to predict, the T1799I substitution is highly likely to destabilize the hairpin conformation. This is consistent with expectations from my hypothesis, where these substitutions increase the relative stability of the loop conformation, yielding a spliceosome with increased capacity.

B. natans – It has two semi-conservative substitutions (F1794A, L1798F) and a non-conservative substitution (T1799A). Both the semi-conservative substitutions do not disrupt any H-bonds within the structure. However, swapping aliphatic and aromatic groups might have other consequences which are hard to predict. On the other hand, T1799, which was described in the previous paragraph, forms a H-bond with its side chain across the hairpin and substitution with A is likely to disrupt this bond and destabilize the hairpin conformation. This is consistent with expectations from my hypothesis where the substitutions increase the relative stability of the loop conformation, yielding a spliceosome with increased capacity.

P. tetraurelia – It has an I1803N substitution. The substitution introduces a polar group in

the side chain in place of an aliphatic, non-polar group. The polar side group of N is highly likely to H-bond with T1789 across the hairpin and stabilize the hairpin conformation. In this case, the substitution increases the relative stability of the hairpin conformation, contrary to expectations from my hypothesis.

Overall, in 2 of the 3 species which use degenerate splice sites, the non-conservative substitutions are predicted to destabilize the hairpin conformation, consistent with expectations from my hypothesis.

Secondary structure predictions for non-conservative substitutions observed in organisms with rigid splice sites suggest an increase in the stability of the hairpin conformation of Prp8 RH extension for some of them, but not all

Of the 12 organisms with rigid splice sites, 9 of them have at least one non-conservative substitution. According to my hypothesis, if conformational stability of Prp8 RH extension is associated with splicing rigid sites, then the substitutions are likely to increase the relative stability of the hairpin conformation. Of the 3 with no non-conservative substitutions, *D. hansenii* is discussed below. *G. theta* (cryptomonad alga) is identical to the hypothetical LECA sequence and *C. parvum* (apicomplexan) has a conservative substitution (T->S). In these two organisms, it is unlikely that RH extension is associated with rigidification. The other species are discussed below.

***E. cuniculi (fungus)* and *C. merolae (red alga)*:** Prp8 in both these organisms is only ~35% identical to human Prp8, the least identity observed in this group of 45 organisms, where the

rest were ~ 50-95% identical to humans. In the multiple sequence alignment, both these organisms are curiously missing most of the RH extension (Figure 3.3). It is hard to predict the effects of missing the RH extension, especially given that many regions across the protein in these two species are different from Prp8 in other organisms. While deleting the extension is lethal in *S. cerevisiae*, it clearly does not have such a drastic effect here in these organisms.

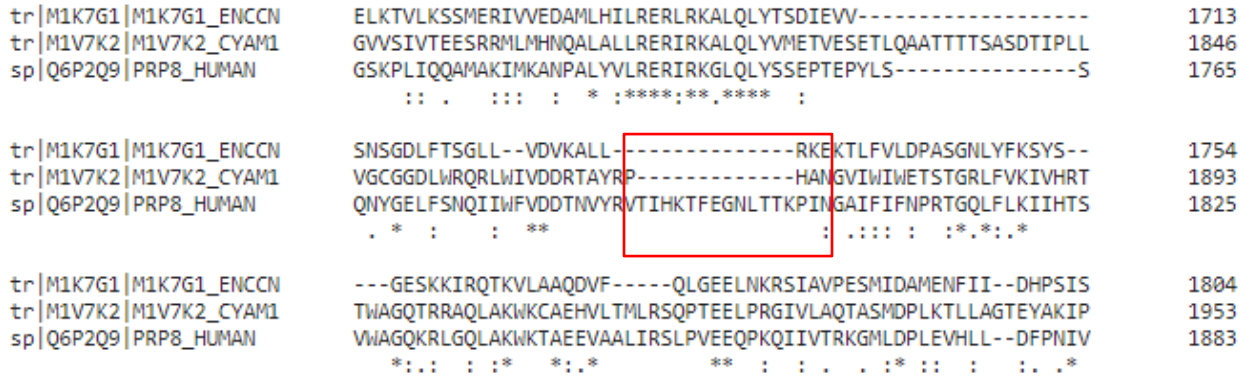


Figure 3.3: A part of the multiple sequence alignment of Prp8 sequence from E. cuniculi, C. merolae and Human. The RH extension region is shown in the red box.
"*" indicates positions of identity, ":" indicates conservative substitutions and "." indicates semi-conservative substitutions.

Trichomonas vaginalis – It has three non-conservative substitutions – H1791Q, T1789K, K1801S. The H1791Q mutation is likely to introduce a H-bond with T1799 across the hairpin. Also, the substitution changes the basic H to an acidic Q group and this is likely to affect other inter-molecular interactions. T1789 is not involved in any interactions through its side chain in the analyzed structure and substitution with K does not disrupt or introduce any

H-bonds. However, it is likely that T1789 might form across hairpin interactions that are not observed in this structure and this will likely be disrupted by the substitution. K1825S mutation is likely to introduce a bond on the same side with P1802 or a bond across the hairpin with T1789, stabilizing the hairpin. Overall, two of the substitutions likely stabilize the hairpin conformation, making the spliceosome more rigid in its selection of splice sites, consistent with expectations from my hypothesis.

Saccharomycetales family: There are 7 species in the Saccharomycetales family investigated in this study and all of them show rigid splice site usage (Bon et al., 2003). Because the time of rigidification in these organisms is unclear, it is unknown whether rigidification happened in the ancestor before they diverged. One possibility is that the ancestor was already rigidified before the 7 species diverged and the other possibility is that all of them evolved rigid site usage independently after diverging. If it is the former case, RH extension might not have an association with rigidification as there is no common non-conservative substitution present in all 7 species which would argue for that substitution being present in the ancestor and being associated with the rigidification event. This is based on the assumption that the observed common conservative substitution does not alter the structure of the hairpin. However, this could be an incorrect assumption and structural studies are needed to verify this. If it is the latter case, then changes in RH extension in each species have to be analyzed independently to see if they might be associated with rigidification.

D. hansenii - Of the 7, *D. hansenii* is the only one to not have any non-conservative substitution. It is unlikely that changes in RH extension were associated with rigidification in this species.

C. albicans, K. lactis, E. gossypi – All these three have the same non-conservative substitution T1799A. This substitution, as described earlier for *B. natans,* destabilizes the hairpin conformation, which is the opposite of what is expected from the hypothesis that increased stability of the hairpin conformation would be associated with rigidification. Therefore, it is unlikely that RH extension influenced rigidificaiton in these species. However, in combination with the conservative substitution that is present, it is unclear what the composite effect would be.

S. cerevisiae, C. glabrata – They both have the same non-conservative substitutions – T1799A and P1802A. The T1799A mutation as described earlier will disrupt the hairpin and this is also observed in the crystal structure of *S. cerevisiae* Prp8 RH extension (PDB – 3e9o)(Pena et al., 2008). The P1803A mutation is an interesting one. Comparing the crystal structures for Prp8 RH extension between human and *S. cerevisiae*, the human extension has only one H-bond between R1787 and N1804 formed by the R side group across the hairpin, whereas the *S. cerevisiae* has two H-bonds at that position. These bonds are likely to stabilize the base of the hairpin. Prolines are known to be rigid and sterically disrupt some secondary structure formations. P1803 residue potentially restricts the positioning of its adjacent residues (including N1804) such that they are not in a suitable position for bonding across with R1787. The difference observed between the human and yeast structures might also just

be an artifact of the crystal formed. Regardless, proline restricts secondary structure formation and therefore, substitution with alanine might make the formation of hairpin conformation easier. Overall, the effects of the two substitutions are hard to predict from structural data and experimental studies on relative stability of hairpin/loop will provide insights.

Yarrowia lipolytica – It has three non-conservative substitutions – T1789A, K1792T and T1800N. The T1789A substitution disrupts an across-hairpin bond that the side chain of T is involved in, thereby likely destabilizing the hairpin. K1816T disrupts bonding of K1816 to 1814 backbone on its own side. The T1800N substitution is likely to introduce an interaction involving the side chain with the side chain of Y1786 across the hairpin, thereby stabilizing the hairpin structure. Overall, it's hard to predict the overall stability of hairpin with one substitution stabilizing and one substitution disrupting it.

With the two caveats mentioned earlier in the section on the inability of this analysis to predict the effects of the composite of all mutations as well as that of conservative substitutions, it is hard to derive concrete inferences from this analysis. Without considering those two aspects, the predictions indicate that half of the organisms have not diverged from the ancestral state in the rigid pool suggesting that changes in RH extension did not drive rigidification in all organisms. However, in the other half of them, RH extension potentially plays a role.

CONCLUSION

Overall, the primary sequence analysis of RH extension correlates strongly with predictions from my hypothesis and shows differential evolution of the extension in rigid and degenerate splicers. The structural analysis shows modest correlation, but it is unclear whether the apparent absence of correlation is a result of (i) the hypothesis being incorrect, or (ii) because our ability to predict stable structures is limited. The current structural predictions are based on a static structure of Prp8. The RNaseH domain of Prp8 shows dramatic movements through the splicing cycle, interacting with different factors. The side chain changes introduced by the substitutions are likely to cause changes in these interactions as well. While one can look at these interactions from structural snapshots of the spliceosome at various points in the pathway, it still does not reflect the dynamic and potentially fleeting nature of these interactions. Therefore, even if the secondary structure predictions provide only a modest correlation, the strong correlation observed in the divergence of the primary sequence is fascinating and warrants further experimental investigation. For instance, swapping the extensions between invariant degenerate and diverged rigid splicers and observing how that affects splice site selection would provide insights into the functional effects of the substitutions in the rigid splicers. One such experiment would be to swap the RH extensions between the degenerate splice site possessing *S. pombe* and the rigid splice site possessing *S. cerevisiae*. Further, while we have investigated only the RH extension here, it would be useful to study other regions of Prp8, including the whole RNaseH domain, and how they have evolved across the degenerate and rigid splicers.

References

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., … Spiegel, F. W. (2012). The Revised Classification of Eukaryotes. *Journal of Eukaryotic Microbiology*, *59*(5), 429–514. https://doi.org/10.1111/j.1550-7408.2012.00644.x

Anantharaman, V., Koonin, E. V, & Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Research*, *30*(7), 1427–1464. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11917006

Bon, E., Casaregola, S., Blandin, G., Llorente, B., Neuvéglise, C., Munsterkotter, M., … Gaillardin, C. (2003). Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Research*, *31*(4), 1121–1135. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12582231

Collins, L., & Penny, D. (2005). Complex Spliceosomal Organization Ancestral to Extant Eukaryotes. *Molecular Biology and Evolution*, *22*(4), 1053–1066. https://doi.org/10.1093/molbev/msi091

Fica, S. M., & Nagai, K. (2017). Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nature Structural & Molecular Biology*, *24*(10), 791–799. https://doi.org/10.1038/nsmb.3463

Galej, W. P., Toor, N., Newman, A. J., & Nagai, K. (2018a). Molecular Mechanism and Evolution of Nuclear Pre-mRNA and Group II Intron Splicing: Insights from Cryo-Electron Microscopy Structures. *Chemical Reviews*, acs.chemrev.7b00499. https://doi.org/10.1021/acs.chemrev.7b00499

Galej, W. P., Toor, N., Newman, A. J., & Nagai, K. (2018b). Molecular Mechanism and Evolution of Nuclear Pre-mRNA and Group II Intron Splicing: Insights from Cryo-Electron Microscopy Structures. *Chemical Reviews*, *118*(8), 4156–4176. https://doi.org/10.1021/acs.chemrev.7b00499

Grainger, R. J., & Beggs, J. D. (2005). Prp8 protein: at the heart of the spliceosome. *RNA (New York, N.Y.)*, *11*(5), 533–557. https://doi.org/10.1261/rna.2220705

Irimia, M., Penny, D., & Roy, S. W. (2007). Coevolution of genomic intron number and splice sites. *Trends in Genetics*, *23*(7), 321–325. https://doi.org/10.1016/J.TIG.2007.04.001

Irimia, M., & Roy, S. W. (2014). Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor Perspectives in Biology*, *6*(6), a016071. https://doi.org/10.1101/cshperspect.a016071

Koodathingal, P., & Staley, J. P. (2013). Splicing fidelity. *RNA Biology*, *10*(7), 1073–1079. https://doi.org/10.4161/rna.25245

Mayerle, M., Raghavan, M., Ledoux, S., Price, A., Stepankiw, N., Hadjivassiliou, H., … Abelson, J. (2017). Structural toggle in the RNaseH domain of Prp8 helps balance splicing fidelity and catalytic efficiency. *Proceedings of the National Academy of Sciences of the*

United States of America, *114*(18), 4739–4744. https://doi.org/10.1073/pnas.1701462114

Pena, V., Rozov, A., Fabrizio, P., Lührmann, R., & Wahl, M. C. (2008). Structure and function of an RNase H domain at the heart of the spliceosome. *The EMBO Journal*, *27*(21), 2929–2940. https://doi.org/10.1038/emboj.2008.209

Plass, M., Agirre, E., Reyes, D., Camara, F., & Eyras, E. (2008). Co-evolution of the branch site and SR proteins in eukaryotes. *Trends in Genetics*, *24*(12), 590–594. https://doi.org/10.1016/J.TIG.2008.10.004

Rogozin, I. B., Carmel, L., Csuros, M., & Koonin, E. V. (2012). Origin and evolution of spliceosomal introns. *Biology Direct*, *7*, 11. https://doi.org/10.1186/1745-6150-7-11

Schellenberg, M. J., Wu, T., Ritchie, D. B., Fica, S., Staley, J. P., Atta, K. A., … MacMillan, A. M. (2013). A conformational switch in PRP8 mediates metal ion coordination that promotes pre-mRNA exon ligation. *Nature Structural & Molecular Biology*, *20*(6), 728–734. https://doi.org/10.1038/nsmb.2556

Schwartz, S. H., Silva, J., Burstein, D., Pupko, T., Eyras, E., & Ast, G. (2008). Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Research*, *18*(1), 88–103. https://doi.org/10.1101/gr.6818908

Sharp, P. A. (2005). The discovery of split genes and RNA splicing. *Trends in Biochemical Sciences*, *30*(6), 279–281. https://doi.org/10.1016/j.tibs.2005.04.002

Will, C. L., & Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, *3*(7). https://doi.org/10.1101/cshperspect.a003707

Yang, K., Zhang, L., Xu, T., Heroux, A., & Zhao, R. (2008). Crystal structure of the beta-finger domain of Prp8 reveals analogy to ribosomal proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(37), 13817–13822. https://doi.org/10.1073/pnas.0805960105

CHAPTER 4

CONCLUDING REMARKS

MECHANISTIC BASIS CONNECTING CONFORMATIONAL TOGGLING IN THE EXTENSION IN THE RNASEH DOMAIN OF PRP8 WITH THE BALANCE BETWEEN SPLICING SPEED AND FIDELITY

Toggle Model at the 2 steps of splicing

Chapter 2 described experiments that test predictions from the Toggle model. Mutations that increase the relative stability of the hairpin/loop conformations of the extension in the RNaseH domain of Prp8 (Prp8 RH extension) were tested for effects on splicing speed and fidelity and the results agree with the model that the hairpin conformation is associated with the transitional state and loop conformation is associated with the catalytic state of the spliceosome.

Our genetic interaction study with Prp2, which activates the spliceosome for the 1st transesterification reaction, showed that increasing the relative stability of the loop conformation suppresses defects in a conditional *prp2* mutant. In contrast, increasing the relative stability of the hairpin conformation shows a synthetic effect. These results are again consistent with expectations from the Toggle model for the first step.

With regards to the second step, while in our study we did not investigate genetic interactions of Prp8 with Prp16, which activates the spliceosome for the second step, observations from a previous study provide some insights. *Liu et. al* looked at genetic interactions of *prp8* 1st step and 2nd step alleles with a *prp16* mutant defective in second step activation. They observed that while the 1st step alleles were synthetic with *prp16*, the 2nd step mutants suppressed the defects (Liu, Query, & Konarska, 2007). As shown in Table 2.1, the classification of *prp8* alleles in the RH extension as 1st and 2nd step mutants based on a

previous 2-step classification (2-step model) (Query & Konarska, 2004) would be equivalent to the transitional/hairpin and catalytic/loop alleles in the Toggle model. While two of the investigated 2nd step alleles in the study have mutations in RH extension and are classified as catalytic/loop alleles in our study, the 1st step alleles investigated are in regions outside of the RH extension and therefore do not provide direct evidence for a synthetic effect for RH extension hairpin/transitional alleles with *prp16*. However, it is highly likely that they will show such a behavior considering that the other 1st step alleles displayed that effect. In conclusion, the above observations are consistent with a model where Prp8's RH extension in the loop conformation is associated with the catalytic state of the spliceosome in the second step as well, while the hairpin conformation is associated with the transitional conformation. While these data demonstrate a correlation between RH extension structure and spliceosome activity, the mechanistic basis by which this correlation is manifested remains unknown. In this chapter, I present a set of hypotheses about how this might occur, and our ongoing experiments designed to better understand this relationship.

How are the loop and hairpin conformations of Prp8's RH extension related to the catalytic and transitional states of the spliceosome?

Model for the involvement of Prp8's RH extension at the 1st step

The activation for the 1st step requires Prp2 activity. Before activation, the active site is formed and the 5'splice site is held there by U6 snRNA. However, the BP-adenosine is still away from the active site, bound by Hsh155 while the branch helix is bound by other SF3 factors (Figure 4.1a) (Fica & Nagai, 2017). Following Prp2 activity, where it binds to the

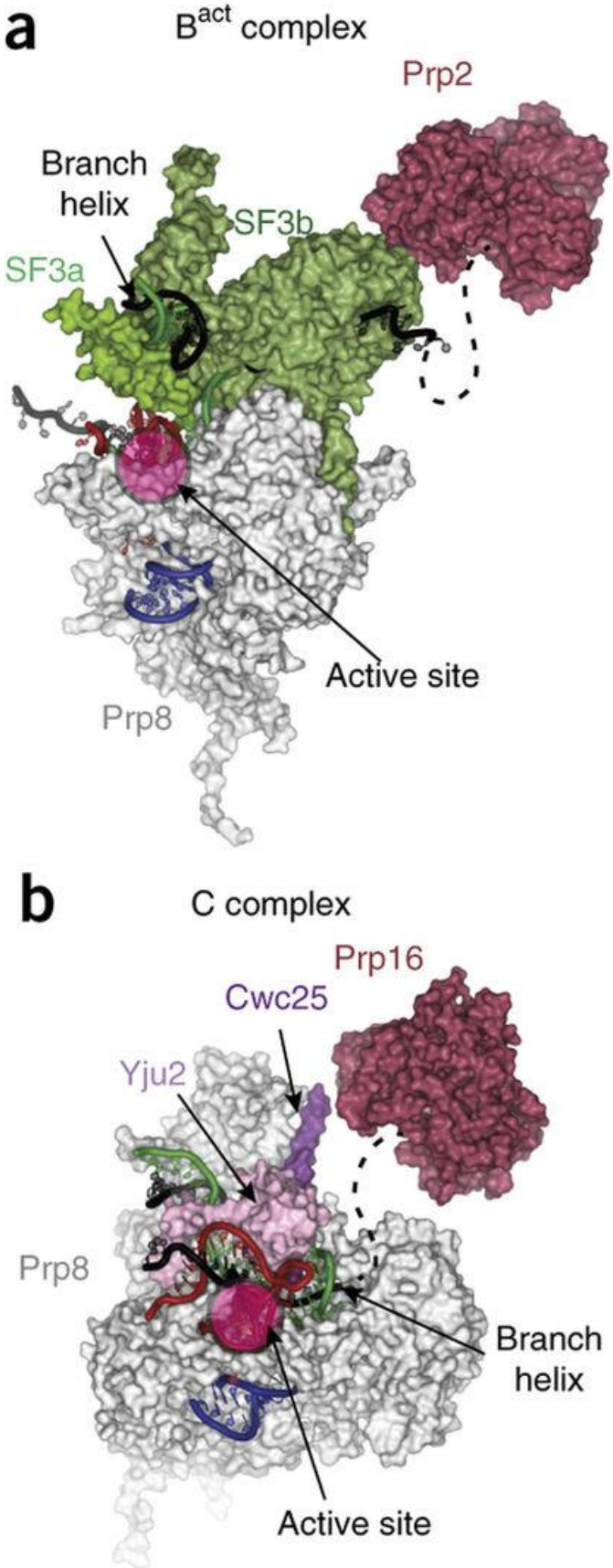
pre-mRNA downstream of the SF3 complex and 'tugs' along it, the SF3 complex is displaced, releasing the BP-adenosine and enabling its entrance into the active site.

Where do Prp8's RNaseH domain and RH extension fit in this? The RNaseH domain sits near Hsh155 which covers the BP-adenosine. The downstream Jab1/MPN domain of Prp8 interacts with Brr2 which is in contact with Prp2 (Fica & Nagai, 2017a). Thus, movement in the RNaseH domain could act in concert with Prp2 to displace the SF3 complex, including Hsh155. The loop form of RH extension is rotated 45^{o} from the position in the hairpin form, and the orientation and exposure of side chains in the extension is altered (Schellenberg et al., 2013). I propose that a switch from the hairpin to the loop conformation of RH extension just around Prp2 activation enables new interactions with nearby surfaces and facilitates movement of the RNaseH domain that in turn helps in the Prp2-mediated displacement of the SF3 complex, including Hsh155.

Not only is the SF3 complex displaced, but Prp2 activity has also been suggested to induce destabilization of the catalytic core. Wlodaver & Staley had shown in 2014 that Prp2 activity destabilizes the RNA elements at the catalytic core - U2/U6 helix Ia, Ib and the catalytic triplex whose residues bind to Mg^{2+} ions involved in catalysis - and then the catalytic core reforms for 1^{st} step catalysis (Figure 1.3) (Wlodaver & Staley, 2014). What might be the significance of disruption of the catalytic site? The disruption induced in U2/U6 helix I by the action of Prp2 presumably provides flexibility and room for movement of the downstream branch helix (U2 snRNA – pre-mRNA BP pairing) and positioning of the BP-adenosine into the active site for the 1^{st} step. Indeed, it was shown recently that Prp2 action provides flexibility to the 1^{st} step reactants. (Bao, Höbartner, Hartmuth, & Lührmann, 2017).

Figure 4.1: Prp2 and Prp16 in B^{act} and C complexes (reproduced from Fica and Nagai, 2017)

Prp2 (a) and Prp16 (b) bind to the pre-mRNA (black) downstream of the branch helix and displace SF3 (a) and first step factors (Yju2, Cwc25) (b) from the intron respectively. The active site rests on Prp8 (white).


a
B^act complex
Prp2
Branch helix
SF3b
SF3a
Active site
Prp8
b
C complex
Prp16
Cwc25
Yju2
Prp8
Branch helix
Active site

How might the RH extension be involved in catalytic core disruption and reformation during 1st step activation? In the C complex, which forms right after the 1st transesterification reaction, the RH extension is observed to bind along a region of the branch helix (Fig 1.6c)(Fica & Nagai, 2017). This state presumably exists before the 1st transesterification reaction helping stabilize the branch helix and positioning the BP-adenosine in the active site for catalysis. Because the branch helix is stabilized in a particular position, this likely helps in the reformation of the catalytic RNA core which was deformed after Prp2 activation for catalysis. I propose that the loop conformation of RH extension provides more flexibility than the hairpin conformation to find and bind along the branch helix and facilitate faster stabilization of reactants and the RNA core in the active site. Indeed, genetic interactions with mutations in U6 snRNA which destabilize U2/U6 helix Ia have shown that the conditional lethality of the U57C mutant is rescued by 2nd step (catalytic/loop alleles) *prp8* alleles (Query & Konarska, 2004). This fits with the model where the loop conformation provides more flexibility to the extension and results in faster binding to the branch helix, thus stabilizing the catalytic core after Prp2 activation, even though the U2/U6 helix Ia is destabilized by the mutation in U6.

I also propose that the loop conformation is transient. The RH extension switches to the loop conformation around the time of Prp2 activity, helping displace the SF3 complex, reform the catalytic core and stabilize the reactants for the 1st step. Then, it switches back to the hairpin conformation following its binding along the branch helix stabilizing it in that state until after the 1st step.

Model for the involvement of Prp8's RNaseH domain and RH extension at the 2nd step

Similar to Prp2, Prp16, acting before the 2nd transesterification, has been suggested to disrupt the catalytic RNA core (Mefford & Staley, 2009) as well as displace the first step factors (Cwc25, Yju1 and Isy1) that are bound to the pre-mRNA (Figure 4.1b)(Fica & Nagai, 2017b). This displacement of first step factors and disruption of catalytic core presumably helps in the movement of branch helix, where the BP-adenosine is now bonded to the 5'splice site in a lariat linkage, out of the active site to allow room for the 3'-exon, the substrate for the 2nd step.

Before Prp16 activity, the RNaseH domain interacts with the first step factors (Cwc25 and Yju1) and the RH extension holds the branch helix. A shift from hairpin to loop conformation in RH extension is likely to again provide flexibility in movement and loosen the hold on the branch helix. Thus, the RNaseH domain and RH extension could act in concert with Prp16 to displace first step factors and get the lariat intermediate out of the active site.

After Prp16's activity, the RH extension extends into the groove of the branch helix and sits at the interface between the branch helix and U6 snRNA region that binds to the 5' splice site (Figure 1.5d). This likely helps in keeping the lariat intermediate out of the active site allowing room for the 3'-exon into the active site and activating the spliceosome for 2nd step catalysis. Also, this might restrict the flexibility of the branch helix/U2 movement and help in the reformation of the catalytic RNA core disrupted by Prp16 activity. The loop conformation of RH extension, with its different orientation and exposed side chains, might make it easier for the RH extension to fit into the groove of the branch helix. The loop conformation might again be a transient conformation to facilitate

these changes, as in the case of the 1st step.

Overall model for RH extension toggle at the 2 steps

In essence, here is my model for how the two conformations of RH extension are associated with the catalytic and transitional states. RH extension in the loop conformation provides more flexibility of movement than the hairpin conformation. In the loop conformation, H-bonds between side chains across the structure might be limited in comparison to the hairpin conformation owing to their non-proximity. This allows more of the side chains to be exposed to the surrounding environment and interact with surrounding partners that allows for flexibility in movement. Owing to the flexibility in movement, in the loop conformation, it is easier to switch positions and facilitate RNP remodeling that activates the spliceosome for catalysis. Therefore, the loop conformation is associated with the catalytic state of the spliceosome. By contrast, the hairpin conformation is less efficient in facilitating RNP remodeling for spliceosomal activation and is therefore associated with the transitional state of the spliceosome.

Studying the kinetics of movement of the RH extension around the active site and whether it is altered in the transitional and catalytic *prp8* alleles would test the prediction about the loop conformation providing more flexibility and in turn faster activation. One can envision such an experiment using FRET with one probe in the RH extension and another tethered to U5 snRNA. U5 snRNA binds the 5'-exon near the active site and stays there from B^{act} until the two chemical steps are done (Fica & Nagai, 2017) and hence, would be a suitable substrate for tagging as it would allow monitoring the kinetics of the movement of the RH extension around the active site through both the steps of splicing.

Kinetics of the movement would provide valuable insights into our understanding of how the two conformations influence the speed of splicing.

What triggers the conversion from one conformation to another during splicing?

This is unclear yet. It could be triggered by movements in the interacting partners, which bring the RH extension in touch with new interfaces that eventually change its conformation. The movement in interacting partners might in turn be tuned by the geometry of the active site. Indeed, one such model that has been proposed is where Cwc25 acts as a transducer between the RNaseH domain and the active site, transmitting information from one end to the other during 1st step activation (Figure 1.6c) (Abelson, 2017; Galej et al., 2016). We still do not have a complete understanding of all the interactions that happen to produce the structural rearrangements. A global screen for genetically interacting partners with the catalytic and transitional *prp8* RH extension alleles would help in identifying the factors that act in concert with the extension at the catalytic steps.

While RH extension was the main focus of this thesis, there might be other factors that toggle between catalytic and transitional states. For instance, in Chapter 2 U2 stem II was mentioned to be part of this toggle. Overall, many components are likely to act in concert to switch the spliceosome between catalytic and transitional states.

CHARACTERIZING THE EFFECTS OF MUTATIONS IN THE RH EXTENSION IN PROCESSING COMPLEX SPLICE SITES USING *S. POMBE* AS A MODEL

The variation in RH extension observed in organisms using rigid splice sites in comparison to those using degenerate splice sites, as observed in the study in Chapter 3, points to the RH extension being a potential influencer in the splice site nature observed in organisms. Yeast has been a great model to understand the effects of the two conformations of RH extension on splicing efficiency/fidelity owing to its genetic tractability and availability of an immense resource of mutants in other splicing factors. However, its limitation arises from a relatively simple splicing setup in comparison to higher eukaryotes like humans. Importantly, yeast with its rigid splice sites does not allow an in-depth understanding of how the optimality of substrates, when the sequences are degenerate, is worked out by the spliceosome. Bridging this gap, the fission yeast *Schizosaccharomyces pombe* would be a suitable model to understand more complex splicing (Fair & Pleiss, 2017). It is genetically tractable like budding yeast, yet possesses some level of complexity in splicing like humans, including degeneracy in splice site sequences.

The data described in Chapter 2 showed that different changes in the sequences of RH extension can have different outcomes on splicing speed and fidelity, depending on the degree to which the relative stabilities of the two conformations are altered. Therefore, it would be useful to map the effects of the various possible mutations that can be made in the RH extension on splicing. This would provide important mechanistic insights about the structural toggle. As a first step towards generating a functional map of RH extension, I have

used a saturation mutagenesis approach using Programmed ALlelic Series (PALS) (Kitzman, Starita, Lo, Fields, & Shendure, 2015) to mutate every single amino acid in the 17 amino acid RH extension to all the other 19 amino acids as well as stop codon in *S. pombe*.

This library can be used to explore the consequences of single mutations in a high-throughput functional screen. One can perform screens for splicing efficiency with splice sites of varying strength and understand how changes in RH extension affect their usage. The expectation is that the more stabilized the loop conformation is, weaker sites will get used with higher efficiency than in a WT strain. In contrast, increasing the stability of the hairpin conformation is expected to decrease the efficiency of usage of weak sites. Further, fidelity is also measurable on a large scale using in vivo native introns showing usage of canonical and aberrant sites (Mayerle et al., 2017). An RT-PCR experiment to measure canonical and aberrant lariat isoforms as described in Figure 2.8B can be done in a high throughput fashion and coupled with deep sequencing using a technology recently developed in the lab (Larson, Fair, & Pleiss, 2016) to quantify aberrant site usage in each mutant. The resulting fidelity measurement for each mutant can be used to draw a sequence-function map.

REFERENCES

Abelson, J. (2017). A close-up look at the spliceosome, at last. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(17), 4288–4293. https://doi.org/10.1073/pnas.1700390114

Bao, P., Höbartner, C., Hartmuth, K., & Lührmann, R. (2017). Yeast Prp2 liberates the 5' splice site and the branch site adenosine for catalysis of pre-mRNA splicing. *RNA (New York, N.Y.)*, *23*(12), 1770–1779. https://doi.org/10.1261/rna.063115.117

Fair, B. J., & Pleiss, J. A. (2017). The power of fission: yeast as a tool for understanding complex splicing. *Current Genetics*, *63*(3), 375–380. https://doi.org/10.1007/s00294-016-0647-6

Fica, S. M., & Nagai, K. (2017a). Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nature Structural & Molecular Biology*, *24*(10), 791–799. https://doi.org/10.1038/nsmb.3463

Fica, S. M., & Nagai, K. (2017b). Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nature Structural & Molecular Biology*, *24*(10), 791–799. https://doi.org/10.1038/nsmb.3463

Galej, W. P., Wilkinson, M. E., Fica, S. M., Oubridge, C., Newman, A. J., & Nagai, K. (2016). Cryo-EM structure of the spliceosome immediately after branching. *Nature*, *537*(7619), 197–201. https://doi.org/10.1038/nature19316

Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., & Shendure, J. (2015). Massively parallel single-amino-acid mutagenesis. *Nature Methods*, *12*(3), 203–206. https://doi.org/10.1038/nmeth.3223

Larson, A., Fair, B. J., & Pleiss, J. A. (2016). Interconnections Between RNA-Processing Pathways Revealed by a Sequencing-Based Genetic Screen for Pre-mRNA Splicing Mutants in Fission Yeast. *G3 (Bethesda, Md.)*, *6*(6), 1513–1523. https://doi.org/10.1534/g3.116.027508

Liu, L., Query, C. C., & Konarska, M. M. (2007). Opposing classes of prp8 alleles modulate the transition between the catalytic steps of pre-mRNA splicing. *Nature Structural & Molecular Biology*, *14*(6), 519–526. https://doi.org/10.1038/nsmb1240

Mayerle, M., Raghavan, M., Ledoux, S., Price, A., Stepankiw, N., Hadjivassiliou, H., … Abelson, J. (2017). Structural toggle in the RNaseH domain of Prp8 helps balance splicing fidelity and catalytic efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(18), 4739–4744. https://doi.org/10.1073/pnas.1701462114

Mefford, M. A., & Staley, J. P. (2009). Evidence that U2/U6 helix I promotes both catalytic steps of pre-mRNA splicing and rearranges in between these steps. *RNA (New York, N.Y.)*, *15*(7), 1386–1397. https://doi.org/10.1261/rna.1582609

Query, C. C., & Konarska, M. M. (2004). Suppression of Multiple Substrate Mutations by Spliceosomal prp8 Alleles Suggests Functional Correlations with Ribosomal Ambiguity Mutants. *Molecular Cell*, *14*(3), 343–354. https://doi.org/10.1016/S1097-2765(04)00217-5

Schellenberg, M. J., Wu, T., Ritchie, D. B., Fica, S., Staley, J. P., Atta, K. A., … MacMillan, A. M. (2013). A conformational switch in PRP8 mediates metal ion coordination that promotes pre-mRNA exon ligation. *Nature Structural & Molecular Biology*, *20*(6), 728–734.

Wlodaver, A. M., & Staley, J. P. (2014). The DExD/H-box ATPase Prp2p destabilizes and proofreads the catalytic RNA core of the spliceosome. *RNA (New York, N.Y.)*, *20*(3), 282–294.

APPENDIX 1

WIDESPREAD ALTERNATIVE AND ABERRANT SPLICING REVEALED BY LARIAT SEQUENCING

Nicholas Stepankiw, Madhura Raghavan, Elizabeth A. Fogarty,

Andrew Grimson & Jeffrey A. Pleiss

Department of Molecular Biology and Genetics
Cornell University, Ithaca NY 14853, USA

This work was published in Nucleic Acids Research, Volume 43, Issue 17, 30 September 2015, Pages 8488–8501

Author Contributions: Experiments were conceived and designed by N.S., M.R., E.A.F., A.G., and J.A.P. Experiments were performed by N.S., M.R., and E.A.F. Data were analyzed by N.S., M.R., A.G., and J.A.P. The manuscript was written by N.S., A.G. and J.A.P.
I was specifically involved in performing qRT-PCR measurements of relative lariat levels in two-intron genes.

ABSTRACT

Alternative splicing is an important and ancient feature of eukaryotic gene structure, the existence of which has likely facilitated eukaryotic proteome expansions. Here, we have used intron lariat sequencing to generate a comprehensive profile of splicing events in *Schizosaccharomyces pombe*, amongst the simplest organisms that possess mammalian-like splice site degeneracy. We reveal an unprecedented level of alternative splicing, including alternative splice site selection for over half of all annotated introns, hundreds of novel exon-skipping events, and thousands of novel introns. Moreover, the frequency of these events is far higher than previous estimates, with alternative splice sites on average activated at ~3% the rate of canonical sites. Although a subset of alternative sites is conserved in related species, implying functional potential, the majority are not detectably conserved. Interestingly, the rate of aberrant splicing is inversely related to expression level, with lowly expressed genes more prone to erroneous splicing. Although we validate many events with RNAseq, the proportion of alternative splicing discovered with lariat sequencing is far greater, a difference we attribute to preferential decay of aberrantly spliced transcripts. Together, these data suggest the spliceosome possesses far lower fidelity than previously appreciated, highlighting the potential contributions of alternative splicing in generating novel gene structures.

INTRODUCTION

The process of pre-messenger RNA (pre-mRNA) splicing, mediated by the spliceosome and accessory proteins, removes non-coding introns, which are found throughout most eukaryotic transcripts (1). Interactions between the spliceosome and sequences within introns are central to the mechanism of splicing. The essential sequences are the 5′ and 3′ splice sites (5′ and 3′SS), found at the termini of the intron, and an internal sequence known as the branchpoint (BP) (2). Along with spliced exons, excised introns are released from the spliceosome as a lariat, in which the 5′SS is covalently attached via a 2′-5′ phosphodiester linkage to an essential adenosine within the BP sequence (3). Thus, the progress and patterns of splicing can be monitored by methods that detect either the mature spliced mRNA or the intron lariats.

In higher eukaryotes, alternative splicing provides a powerful opportunity for both regulation of gene expression and generation of proteome diversity (4). Mechanistically, alternative splicing derives from control of splice site selection (5). The splice site sequences in mammalian introns are highly degenerate, often requiring the activity of auxiliary proteins to enhance activation of sub-optimal sequences (6). Because of the low information content at many mammalian splice sites, nearby cryptic sites can be activated, resulting in the production of alternative splice products, many of which are subjected to degradation by RNA quality control pathways (7). Several estimates of the intrinsic rate at which the spliceosome aberrantly generates alternative splicing products have been previously published. A quantitative RT-PCR based study of mammalian splicing suggested an exceptionally low rate of aberrant exon skipping, on the order of one error for every 103–105splicing events (8). By contrast, computational modeling of exon skipping and alternative

splice site usage within mammalian EST data predicted higher error rates that were variable, dependent upon expression level and intron number of the host transcript (9). And finally, analyses of large RNA-seq datasets have identified alternative events at a rate of one in 102–103 splicing events (10). Importantly, it is unclear whether any of these approaches are appropriate for defining the actual spliceosomal error rate given that they predominantly measure mature mRNA levels, which derive from both rates of splicing error and preferential decay of mispliced products.

Here we examine global splice site selection in the fission yeast, Schizosaccaromyces pombe. Unlike budding yeast, splice site sequences in S. pombe are highly degenerate and comparable to those found in mammalian introns (11,12). Furthermore, nearly half of all S. pombe genes contain an intron, and nearly half of those contain multiple introns. We previously described a lariat sequencing approach that allows for high depth analysis of splicing events (13). Here, we have improved upon this approach to generate the most comprehensive dataset of experimentally identified splicing events in an organism with degenerate splice site sequences. We observe a large number of alternatively spliced isoforms, a subset of which corresponds to conserved alternative splice sites. Remarkably, we also observe a rate of alternate splice site selection that is greatly higher than previous estimates. Together, these data provide compelling evidence suggesting that spliceosomal infidelity is widespread, and provides a powerful mechanism by which alternative gene structures can evolve (14).

MATERIALS AND METHODS

Strains used

yNZS005 was used as the WT strain and was from ATCC (Linder 972, #38366). yNZS006 contained a deletion of the dbr1 locus from yNS005 and was generated as previously described (13). yNS008 contained a deletion of the upf1 locus, and was taken from the S. pombe Haploid Deletion Collection from Bioneer, and yNS007 was the matched wild type strain from this library.

Yeast cultures

Unless otherwise noted, yeast cultures were grown according to standard protocols (15). Cells were harvested by filtration thru Millipore HAWP0025 filters when OD600measurements were between 0.8 and 1.0. For heat shock samples, after reaching the noted OD600 the cultures were shifted to 37°C for 15 minutes and then harvested. For the diauxic shift samples, the cultures were allowed to grow until the OD600 reached 7.6.

Two-dimensional (2D) gel electrophoresis

Denaturing acrylamide gels were polymerized with 7.5 M Urea with varying acrylamide concentrations. Two sets of gels were run: one designed to isolate shorter lariats and the other to isolate longer lariats. For shorter lariats, the acrylamide percentages were 7.5% and 15% for the first and second dimensions, respectively. For the longer lariats, the percentages were 4% and 8%. For both gel types, 40–60 μg of Δdbr1 RNA was loaded onto the first gel. The first gel was run at 200 V for 2 h and the second gel at 230 V for 2 h followed by 290 V for 1 h. The entire lane containing the RNA was cut from the gel, rotated, and recast at the top of the second gel. The second gels were run until the xylene cyanol dye migrated 5.5 or 9 cm for the short and long lariats, respectively. Gels were stained, visualized (Dark Reader or

Typhoon) and lariat arcs were excised. RNA was extracted according to standard protocols.

Lariat sequencing library construction

The RNAs from 2D gels were used to construct non-stranded multiplexed libraries using a custom protocol that placed appropriate sequences for multiplexed Illumina-sequencing. The first strand and second strand synthesis was done using Invitrogen second strand synthesis kit and 500 ng of dN9 primer. Because of the high levels of dN9 present in our first strand reaction, we omitted the Escherichiacoli DNA ligase from the second strand reaction. After second strand synthesis, the products were run on a native gel and the library was recovered between ∼20 and 300 nucleotides. The sized material was eluted from the gel by adding 4x volume of 0.3 M NaOAc pH 5.3. The DNA was ethanol precipitated by adding 2.5 volumes of 95% ethanol, incubating at −20°C, spinning at 14 000 × g for 20 min, washing twice with 70% ethanol for 10 min and resuspending in 10 μl H2O. For all subsequent steps, enzymatic reactions were purified using phenol:chloroform extraction followed by ethanol precipitation. DNA end repair was performed using NEB Next End Repair Module. Next, dA tailing was performed using the NEB Next dA-tailing Module. Adapter ligation was performed using T4 DNA Ligase (Rapid) from Enzymatics and 1 μl of Illumina barcoded adapter. The resulting ligation product was sized on a denaturing urea gel between 130 and 300 nucleotides. This was precipitated and sequenced on an Illumina Hiseq 2000 with single-end 100 nucleotide reads.

Lariat sequencing genome alignment

Illumina sequence reads were trimmed of the 3′ adapter using Trimmomatic (16) with parameters:3:30:10 and MINLEN:18. Trimmed reads were aligned to the S. pombe genome (pombase.org, Schizosaccharomyces_pombe.ASM294v2.21.dna.genome.fa) using Bowtie2

(17). Parameters of alignment were ‘score-min L,-0.4,-0.4 –very-sensitive’. End-to-end alignments were done for the initial genome alignment. Paired end alignments on split reads were done using ‘–ff -I 20 -X 3000 –no-mixed –no-discordant’.

Splice site scoring

Log-odds scores for splice sites were computed from a Position Weight Matrix (PWM) (6). The 5′SS was scored either using the dinucleotides within the first nine nucleotides of annotated introns or using the dinucleotides within three nucleotides upstream the 5′SS through the first nine nucleotides of annotated introns for the foreground signal and the dinucleotides of annotated intron sequences for the background signal. The former was used during splitting reads only, while the later was used for all analysis after split read identification. The putative BP was scored using an 8 position dinucleotide model using branch points identified from the aligned_introns.txt file from pombase.org for the foreground signal and the dinucleotide composition of introns for background signal. The BP and 9-nucleotide 5′SS scores have a similar maximal value in this scheme.

Branch spanning split read alignment

For identification of branch spanning reads, each read that failed to align to the genome in end-to-end fashion was split into paired-end reads at every GT dinucleotide, including those that appeared in the reverse complement of the strand. For alignment, we only considered the subset of these reads for which the splitting process produced two fragments of at least 10 nucleotides, and excluded all others. These reads were then assessed for alignment using Bowtie2 in paired-end mode with the fragment containing the GT as mate 1 and the other as mate 2, and using the previously noted parameters. Since a single read split in this fashion

can yield several possible alignments, possible GT split alignments were collapsed into a best available paired-end alignment that minimizes the number of mismatches in the alignment. As an additional criterion to judge alignment quality, log-odds scores were calculated for the putative 5′SS (GT end of the alignment) and BP (the other end of the alignment): the combination of these scores was required to be greater than 0; this strategy was used to both break ties and reduce the number of artifactual alignments.

Reverse transcriptase frequently introduces deletions and mutations when creating the cDNA product that crosses the 5′SS to the BP (18). To determine the likely branch point nucleotide, BP scores were computed between two nucleotides upstream of the non-GT end of the paired alignment and three nucleotides downstream of the read end and the position of the BP was determined by the maximal BP score in that range. Due to the to the high rate of mismatches from reverse transcriptase reading across the branch point of the lariat, up to two mismatches within ±1 nucleotide of the putative branchpoint adenosine were disregarded for total mismatch calculations. The position where the BP minimized the number of mismatches was considered the branch point. If the branch point score caused the total log-odds scores to fall below 0, then the read was considered to fail alignment.

At a low frequency, the heuristics of Bowtie2 are such that it can fail to find an alignment when an acceptable alignment is possible. At a low rate this results in incorrect alignments where a nearly correct version of the GT iteration with some mismatches has an alignment but the correct alignment failed to align. Found alignments were assessed for better nearby alignments by appropriately shifting sequence from one end of the alignment to the other side. This shifted sequence was checked to see if it decreased the number of total mismatches

in the alignment. The BP was then reevaluated in this alignment, as described above. Additionally, since Bowtie2 may have failed to correctly align a read to the genome in end-to-end fashion, found split read alignments were assayed for this by anchoring one end of the mate-pair to the genome and checking if recreating the original read leads to the non-anchored mate aligning using the Smith–Waterman algorithm. If this original-format alignment was plausible, then the paired-end read alignment was removed from further analysis.

To estimate an upper bound for false alignments generated by our split-read approach, the alignment strategy described above was applied to the reads that aligned to the genome with one difference: the initial reads were split at GA dinucleotides instead of GT. Of the ~78 million genome-aligning reads assessed this way, only 15 727 reads could be split and aligned to the genome, reflecting a total of 1267 intron branch intervals, yielding a false alignment rate of only ~0.02%. Importantly, this rate may well be an over estimate because of the propensity of some 'true' split reads to incorrectly align to the genome with mismatches.

Branch read identification

Strandedness of a split read was determined by selecting the highest scoring 5′SS and BP scores for each direction of the read alignment. The branch point was identified by looking within ±2 nucleotides of the end of the read alignment for the best scoring BP. This allows for small deletions and insertions (a common property of reverse transcription across branchpoints). Split reads that make use of the same 5′SS and BP were aggregated together

and considered to come from the same lariat. The associated 3′SS was determined to be the first AG dinucleotide found at least 5 nucleotides downstream of the BP.

Aggregation of introns

Overlapping split reads were aggregated by strand and by overlapping genomic coordinates. Annotated 5′SS and 3′SS were determined by the POMBASE Schizosaccharomyces_pombe.ASM294v2.21.gff3 file. Alternate sites were defined as any site that did not correspond to either the 5′SS or 3′SS of an annotated intron. Split reads that overlap annotated introns with both splice sites corresponding to a non-annotated location were not further considered. Alternate 5′SS within 5 nt of an annotated 5′SS are prone to mismapping and as an aggregate were ignored as a parsimonious approach to identifying unannotated splicing events. Novel introns were defined as split reads that do not overlap a known intron on the same strand. Exon skipping events were defined as split reads that overlap two introns in the same transcript. Annotated, alternate, and novel introns were required to have at least one read with a minimum of 30 nucleotides of total aligned sequence.

Alternate intron identification

Branches overlapping a single intron were aggregated. For each branch the first AG dinucleotide at least five nucleotides downstream of the intron was called as the 3′SS. For the RNAseq alternate introns, utilization was measured as alternate counts divided by the sum of alternate and annotated counts. When indicated, likelihood of alternate introns were

computed with a binomial distribution created from the number of total reads alternate and annotated reads and an error rate being tested in the range 10^{-1} to 10^{-6}.

RNAseq

RNAseq was done using TruSeq RNA Kit v2 or NEBNext Ultra Directional Kit. RNAseq alignments were done using Tophat suite version 2.0.9 with Schizosaccharomyces_pombe.ASM294v2.21.gff3 and with novel intron discovery. For novel introns, only GT-AG novel exon-exon junctions were further processed. Novel introns found to overlap a single annotated intron were aggregated. Reads for novel intron RNAseq junctions were required to have mapq scores of at least 30. Expression quartiles were computed for transcripts using RNAseq for the indicated RNAseq library by using the exon-exon junction counts from Tophat.

Comparative analyses

MAF format MultiZ alignments(19) from 01/04/2012 were acquired from the Broad Institute website and used to retrieve aligned sequences for computing log-odds scores. Log-odds scores for 5′SSs and BPs were computed, when possible, for each of the *Schizosaccharomyces* species considered. To calculate a background rate of sequence conservation, putative upstream 5′SS and downstream BP sequences were identified from the coding sequences of intronless genes in *S. pombe*. From these sequences, a large number of putative splice sites were initially chosen, the total number representing a ~10-fold increase over the number of identified alternative sites. For each of these sites, the PWM score was determined for the *S. pombe*sequence, after which time a subset of these sites was selected

that had a similar score distribution to the identified alternative sites. For this subset of sites, the PWM was then determined for the orthologous sequence in each of the *Schizosaccharomyces* species, and this score was assessed for conservation. A similar approach was used to determine background rates for the downstream 5′SS and upstream BPs, but using sequences found in *S. pombe* 3′ UTRs that are shorter than 150nt as the source for the background distribution.

qPCR of lariat introns

Intron qPCR measurements were made for two different introns from each of four different multi-intronic genes using RNA isolated from a *Δdbr1* strain. Standard dilution curves using genomic DNA were generated for each primer pair, allowing for comparison of the relative levels of each RNA.

Weblogos

Web logos were generated using a command line version 3.4 (20).

Accession codes

All sequencing data have been submitted at NCBI's Gene Expression Omnibus (GEO) repository with accession number GSE68345

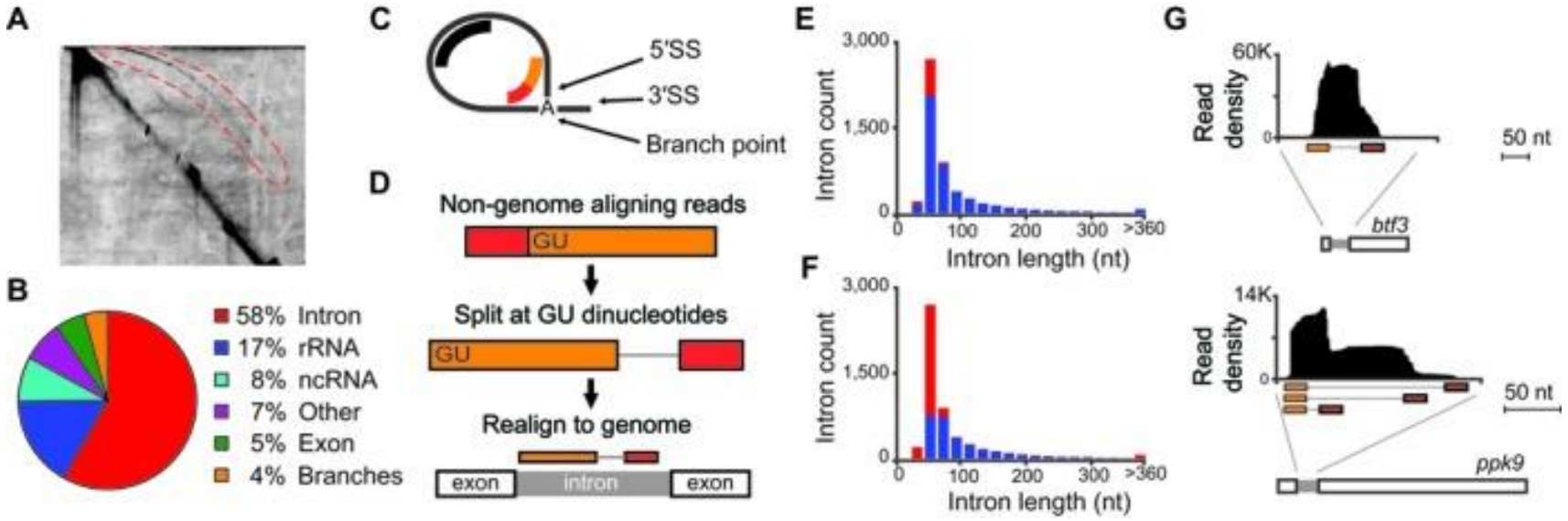
RESULTS

Global *S. pombe* splicing profiles revealed by intron lariat-sequencing

To capture the global splicing profile of *S. pombe*, we used two-dimension gels (FigureA1.1A) to isolate intron lariats from a *Δdbr1* strain grown under several growth conditions (see Materials and Methods). Building upon our previous work (13), experimental conditions were optimized to recover both long and short introns. Purified lariat RNAs were converted, without debranching, into cDNA and sequenced on an Illumina HiSeq 2000, generating over 231 million sequencing reads. Alignment of these reads to the *S. pombe* genome revealed that ~60% of genome-matching reads mapped to annotated introns (Figure A1.1B and Supplementary Table S1) with only a minority of reads mapping to exons, confirming the high level of lariat enrichment afforded by this approach.

Figure A1.1 Intron lariat sequencing defines splicing patterns.

(**A**) Image of two-dimensional gel electrophoresis of RNA isolated from *Δdbr1 S. pombe*. Intron lariats (red-bounded region) were isolated and used as source material for sequencing. (**B**) Pie-chart summarizing allocation of lariat sequencing reads to indicated genomic regions. (**C**) Illustration of intron lariat and splice sites, depicting intron-mapping reads (in black) and branch-spanning reads (in red-orange). (**D**) Schematic of alignment strategy for candidate branch-spanning reads, together with illustration of aligned branch-spanning read. (**E**) Histogram of annotated introns counts (y-axis) separated by length (x-axis, 20 nucleotide bins), indicating introns recovered with intron-mapping reads (blue) and those not recovered (red). (**F**) Histogram of annotated introns counts (y-axis) separated by length (x-axis, 20 nucleotide bins), indicating introns precisely recovered with branch-spanning reads (blue) and those not recovered (red). (**G**) Intron-mapping reads (y-axis indicates read density) aligning (x-axis indicates alignment position) to indicated *S. pombe* pre-mRNAs, together with branch-spanning reads (orange-red) aligned with split-read mapping strategy. The *btf3* peak density truncated at +/-50 nt of intron boundaries. The *ppk9* peak tapers just upstream of intron boundaries.


A
B
58% Intron
17% rRNA
8% ncRNA
7% Other
5% Exon
4% Branches
C
5'SS
3'SS
A
Branch point
D
Non-genome aligning reads
GU
Split at GU dinucleotides
GU
Realign to genome
exon
intron
exon
E
Intron count
3,000
1,500
0
0
100
200
300
>360
Intron length (nt)
F
Intron count
3,000
1,500
0
0
100
200
300
>360
Intron length (nt)
G
Read density
60K
0
50 nt
btf3
Read density
14K
0
50 nt
ppk9

As we and others have previously noted (13,18,21,22), sequencing of lariat introns generates two distinct types of reads: one derived from the body of the intron, and the other derived from reverse transcription across the lariat branch (Figure A1.1C). Branch-spanning reads contain both the 5′SS and BP sequence within the lariat, and are thus diagnostic of a splicing event, somewhat analogous to exon-exon spanning reads within RNAseq data. While these reads are information rich, their identification is non-trivial because of both their inverted nature and the poor efficiency and reduced fidelity of reverse transcription across this junction (18). To systematically identify these reads in our dataset, we developed an alignment pipeline whereby all reads that failed to directly map to the genome were divided at each GU dinucleotide, corresponding to possible 5′SSs, and the pairs of divided reads were re-assessed for alignment to the genome using a split-read mapping strategy (Figure A1. 1D). A total of 3.7 million such reads were identified by this approach with a low false discovery rate (see Materials and Methods, and Supplementary Table S2), making this by far the largest dataset of experimentally identified branch sites to date.

Importantly, the data generated from both the body-mapping and branch-spanning reads successfully identified the majority of known *S. pombe* introns. Over 85% of annotated introns had reads mapped to the body of the intron (Figure A1.1E) and Supplementary Table S3), while ~55% had branch-spanning reads that recover both the annotated 5′SS and predicted BP (Figure A1.1F and Supplementary Table S3). As we and others have previously seen, short intron lariats were particularly difficult to recover in these experiments (13,21). Nevertheless, because branch-reads contain the coupled information of both the 5′SS and BP sequence used to form the lariat, alternative splicing events that would be difficult to reliably

predict from body-mapping reads can be definitively assigned by branch-reads. For example, whereas the peak of body-mapping reads for the *btf3* transcript (Figure A1.1G, top) suggested a discreet 5′SS and BP, the spectrum for the *ppk9* transcript (Figure A1.1G, bottom) suggested possible alternative splice sites. The use of branch-spanning reads readily resolved these different patterns by identifying a single 5′SS/BP combination for *btf3*, but three distinct combinations for *ppk9*. Because of the precision with which branch-reads define splicing events, we relied exclusively upon them for further analysis.

Lariat sequencing identifies widespread examples of alternative splicing

To characterize identified splice sites, a position weight matrix (PWM) scoring metric (6) was implemented (see Materials and Methods), based upon the ~5,000 splice sites annotated on PomBase (23). Previous computational analyses had predicted the likely BP for >90% of annotated *S. pombe* introns; these sites were used as the basis of our PWM scoring for BPs. For the small number of introns for which multiple BPs were predicted, we considered the BP closest to the annotated 3′SS to be the primary BP, and used it for our analysis. As expected, our analysis of the set of annotated introns showed a wide range of scores for both the 5′SS and BP sequences, reflecting the degeneracy of splice site sequences in *S. pombe* (Figure A1.2A) (11). Interestingly, while the scores of the 5′SSs recovered by lariat sequencing closely match the distribution of all annotated scores, the recovered BPs included many more low scoring sequences. Remarkably, nearly 900 of the annotated introns recovered here revealed activation of multiple BPs, each of which was predicted to use the annotated 3′SS. Importantly, while this estimate represents the lower limit of the frequency of alternate BP activation because of the size limitations described previously, it is

nevertheless significantly higher than a previous study indicated (21). When considering the scores of the primary BPs we identified, defined as the closest BP upstream of the annotated 3′SS, there was little difference between those annotated introns for which only a single BP was identified and those with multiple branches (Figure A1.2B), suggesting that alternative BP selection is not driven simply by the strength of the primary BP. Not surprisingly, however, quality scores of alternative BPs tended to be weaker than those of primary branch sites (Figure A1.2B). Nevertheless, many annotated introns contain alternative BP sequences with scores comparable to those of the presumed primary branch sites; representative examples are shown for the *spf47* intron 1 (Figure A1.2B) for which the identified alternative and presumed primary branch points both have scores near the middle of their respective distributions.

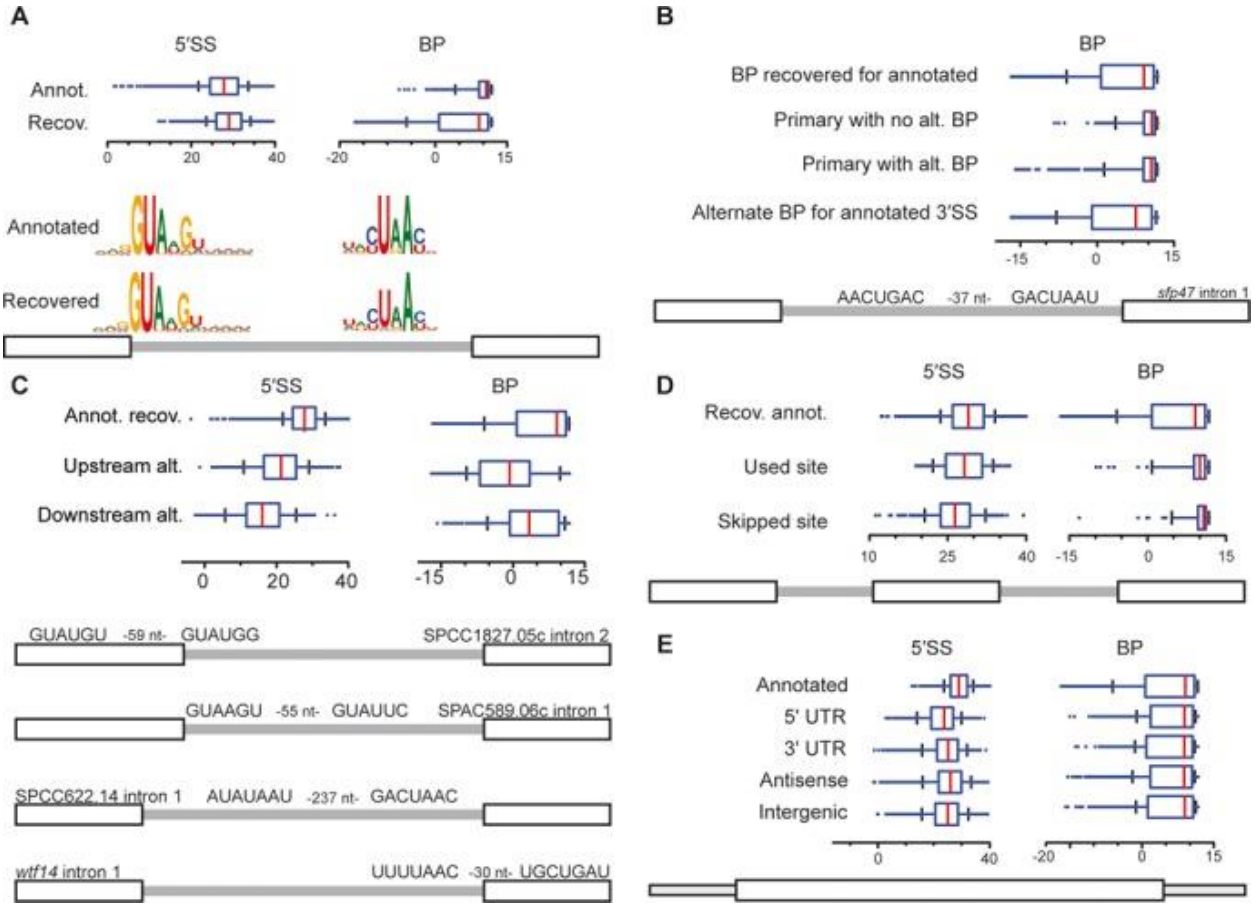


Figure A1.2

Global analysis of alternative and novel splice sites in *S. pombe*. (**A**–**E**) Distribution of splice-site strengths as boxplots (x-axis indicates PWM scores), together with examples of alternative site sequences. (**A**) Splice site scores (5′SS and BP) corresponding to annotated introns (annot.), and sites corresponding to annotated introns recovered with branch-spanning reads (recov.) (**B**) BP scores for indicated categories of alternative splicing events associated with alternative BPs that are paired with annotated 3′SSs. (**C**) Splice site scores (5′SS and BP) corresponding to alternative splice site scores partitioned into upstream and downstream alternative intron boundary sites, compared to annotated and recovered sites (annot. recov.). (**D**) Splice site scores (5′SS and BP) corresponding to splice site scores associated with exon-skipping events partitioned into sites participating in exon-skipping (used site) and those skipped (skipped site), compared to annotated and recovered sites (annot. recov.). (**E**) Splice site scores (5′SS and BP) corresponding to sites found in novel introns in indicated genomic regions, compared to annotated and recovered sites (annot. recov.).

Whereas the alternative branch points noted above are not predicted to change intron-exon boundaries, an additional 2,923 alternative splicing events (associated with 1851 annotated introns) were identified that utilize one annotated site and one alternate site and are predicted to change the coding character of the resulting mRNA. Remarkably, these alternative splicing events implicate at least half of all annotated introns as subject to alternative splicing. Included among these were 1031 events (corresponding to 858 annotated introns) where an alternative 5′SS is spliced to the canonical BP, and 1892 events (corresponding to 1276 annotated introns) where the canonical 5′SS is spliced to an alternative BP/3′SS; akin to mammalian alternative splicing. Interestingly, for both 5′SSs and BPs, alternative sites upstream of the canonical site were identified at nearly twice the frequency as they were downstream (Supplementary Table S4), consistent with a 'first come, first served' model of splice site selection (24).

Many of the alternative splice sites described above had sequence scores similar to those of annotated introns, however, we observed a clear relationship between the quality of the splice sites and their position relative to the annotated sites. For 5′SSs, alternative sites identified upstream of the annotated site had a distribution of scores that, while weaker, substantially overlapped those of the annotated sites (Figure A1.2C). In contrast, the distribution of scores corresponding to alternative downstream 5′SSs were significantly weaker than those of both the annotated and upstream 5′SS (Figure A1.2C). The pattern for alternate BPs was inverted: activated downstream sites had scores more similar to canonical BPs while those found upstream tended to be lower in strength (Figure A1.2C). Representative examples of these types of alternative splicing are shown (Figure A1.2C)

where both the alternative and canonical events have scores near the middle of their respective distributions.

Having identified alternative 5′SSs upstream of nearly 15% of annotated introns, we wondered whether the failure to identify alternative sites for the remaining 85% of introns reflected the absence of an effective alternative splice site or the failure to utilize such sites. To address this question, we examined a 150 nucleotide window upstream of every annotated intron and identified the highest scoring potential 5′SS. Many annotated introns are flanked by upstream sequences that contain high scoring candidate-alternative 5′SSs but for which we detected no alternative splicing (Supplementary Figure S1). Perhaps deeper sequencing might reveal usage of these potential sites; alternatively, additional sequence elements or secondary structures may be functioning to preclude alternative splicing at these locations (25,26).

Previous studies identified exon skipping in *S. pombe*; although relatively few such examples were discovered in earlier experiments (13,21), more recent work identified over 100 high-confidence events (27). Here, we found hundreds of additional exon skipping events, expanding the repertoire of confirmed exon skipping events in *S. pombe* (Supplementary Table S5). Analyses of the splice site sequences associated with the events identified here (Figure A1.2D) revealed that the 5′SS of the upstream intron and the BP sequence of the downstream intron had sequences that were nearly indistinguishable from the composite scores of annotated introns. Remarkably, however, the skipped BPs of the upstream introns were not characterized by low information sequences, but rather appeared to have slightly stronger scores in aggregate than annotated introns, inconsistent with

expectations of intron-definition based models of exon skipping. By contrast, the skipped 5′SS of the downstream introns had significantly weaker splice site scores than annotated introns. This increased propensity of exon skipping events to be associated with weak downstream 5′SSs, together with the absence of weak BP sequences, implies that spliceosome assembly via exon definition may be a more prominent aspect of splicing in *S. pombe* than previously appreciated (28).

In addition to alternative splicing associated with annotated introns, our data also revealed an unprecedented level of splicing across the transcriptome at sites with no characterized introns. A total of 8113 splicing events were identified associated with 7412 novel introns (Supplementary Table S6). These introns were located within the transcripts of protein-coding genes (including 857 within annotated 5′UTRs, 971 within annotated 3′UTRs, and 1567 within the coding regions of these transcripts), within anti-sense RNAs (2699), within non-coding RNAs (554), and within intergenic regions of the genome (1343). Remarkably, while the overall distribution of 5′SSs scores for these novel introns was noticeably lower than those associated with annotated introns, the majority of these novel events had 5′SSs sequences with strong PWM scores (Figure A1.2E). Similarly, the distribution of BP scores for novel introns was virtually indistinguishable from those found in annotated introns (Figure A1.2E). Importantly, although many novel events are recovered with low read counts, the splice-site score distributions are similar across high and low read counts (Supplementary Figure S2).

RNAseq validates widespread alternative splicing

Having identified widespread examples of alternative splicing via lariat sequencing, we turned to RNA sequencing as an orthogonal approach for validation. Datasets of poly(A)$^+$ RNA were generated for both wild-type and *Δdbr1* strains (Supplementary Table S7). Importantly, although loss of Dbr1 perturbed transcript levels of a small subset of genes, the overall transcriptomes of wild-type and *Δdbr1* strains as determined by RNAseq were extremely similar (Pearson correlation coefficient of 0.993, $P < 2.2 \times 10^{-16}$; Supplementary Figure S3). We used TopHat2 to identify transcripts harboring alternative 5′SSs (29), and then examined the extent to which these alternative 5′SSs overlapped with those identified by lariat sequencing. Approximately 40% of alternative 5′SSs identified by lariat sequencing were detectable by RNAseq, and ~25% of alternative 5′SSs identified by RNAseq were detected by lariat sequencing (Figure A1.3A and Supplementary Table S4), demonstrating that the two approaches provide complementary but not identical descriptions of the transcriptome. Importantly, the average quality of alternative 5′SSs, as judged using PWM scoring, was nearly identical when comparing alternative sites defined uniquely by lariat sequencing or RNAseq, whereas those sites identified by both methods generally corresponded to slightly stronger sites (Figure A1.3B). Because lariat sequencing directly identifies BPs but not 3′SS, and RNAseq identifies the inverse, to enable comparison of these datasets the first AG dinucleotide downstream of the BP identified by lariat sequencing was assumed to be the 3′SS. Using this approach, similar overlaps were also observed in the alternative BPs/3′SSs identified by lariat sequencing and RNAseq (Supplementary Figure S4).

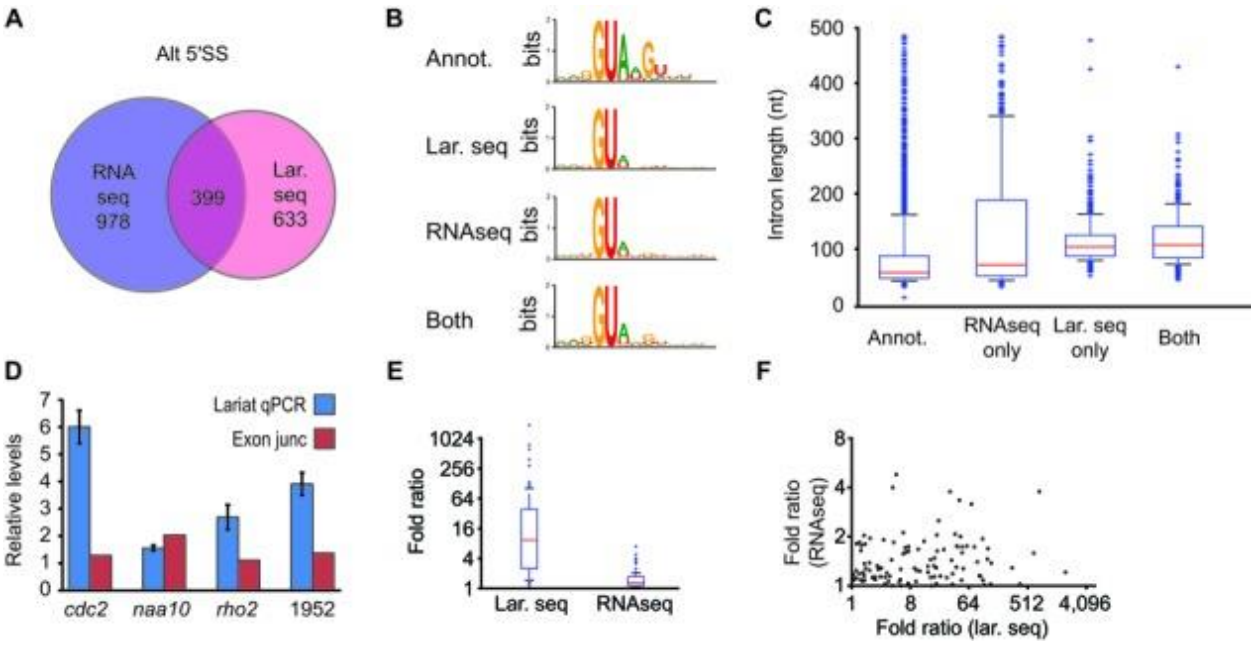


Figure A1.3 Cross validation and comparisons of alternative splicing detected using RNAseq and lariat sequencing.

(**A**) Venn diagram illustrating alternative 5′SSs identified by RNAseq, lariat sequencing, or both. (**B**) Web-logo comparisons of annotated 5′SSs compared to those identified by RNAseq, lariat sequencing, or both. (**C**) Intron length comparisons of annotated introns compared to those identified by RNAseq, lariat sequencing, or both. (**D**) qRT-PCR measurements of relative lariat levels compared for pairs of lariats from two-intron genes (blue), compared to RNAseq determinations of exon-exon junction reads for the corresponding splice junctions (SPAC1952.04c labeled as 1952). (**E**) Boxplots indicating distributions of fold ratios (y-axis) of branch-spanning read counts for pairs of introns from multi-intron genes, compared to ratios of exon-exon spanning read counts for splice junctions from multi-intronic genes with comparably sized intron lengths. (**F**) Scatter-plot of values shown in (e), relating RNAseq-derived ratios (y-axis) to lariat sequencing-derived ratios (x-axis) for genes whose introns are of comparable size.

Multiple sources, both biological and technical, almost certainly contributed to the imperfect overlap between RNAseq and lariat sequencing-based definitions of the transcriptome. A major biological difference derives from the species sequenced: lariats versus mature transcript. This difference is likely to be particularly important for alternative products whose structures result in accelerated decay, leaving them poorly detected by RNAseq. Alternatively, because of the length bias of lariat sequencing, RNAseq is better positioned to capture very short or very long introns, as confirmed in Figure A1.3C. While investigating this length bias, however, we found an additional, and somewhat surprising, result: the levels of lariats derived from different introns of a common pre-mRNA were detected to markedly different degrees in our data, and this difference was maintained even when looking at intron pairs that were of the same general size in genes with multiple introns. Importantly, we confirmed this result using qRT-PCR on four transcripts (Figure A1.3D) by comparing two introns within each transcript. For each mRNA tested, RNAseq indicated that signal derived from the different exon-exon boundaries were narrowly distributed, as expected. In contrast, qRT-PCR measurements indicated that lariat introns derived from the same pre-mRNA were present at greatly different levels (Figure A1.3D). This result was confirmed genome-wide, using exon-exon spanning reads found in RNAseq and branch-spanning reads from lariat sequencing (Figure A1.3Eand F). The biological basis for this result is unclear, but likely indicates variable decay rates for lariats in the absence of Dbr1. Importantly, while this result complicates quantitative comparisons for individual species detected using branch-spanning reads, genome-wide comparisons are less likely to be compromised.

In addition to evidence of extensive alternative splicing in *S. pombe*, our lariat sequencing data also revealed many thousands of novel introns (Figure A1.2E). As before, we wished to validate and compare novel introns found with lariat sequencing to those present in RNAseq. Similar to our previous findings, approximately 15% of novel introns found by lariat sequencing are also found within RNAseq, whereas ~20% of novel introns detected by RNAseq are also found within lariat sequencing (Figure A1.4A and Supplementary Table S8). Importantly, the strengths of 5′SSs and BPs detected by both approaches were highly similar (Figure A1.4B and 4C, respectively), and the lengths of the novel introns recovered by the two approaches are consistent with the previously discussed length biases (Figure A1.4D). Taken together, these results imply the existence of many thousands of additional introns in *S. pombe* not found by either result.

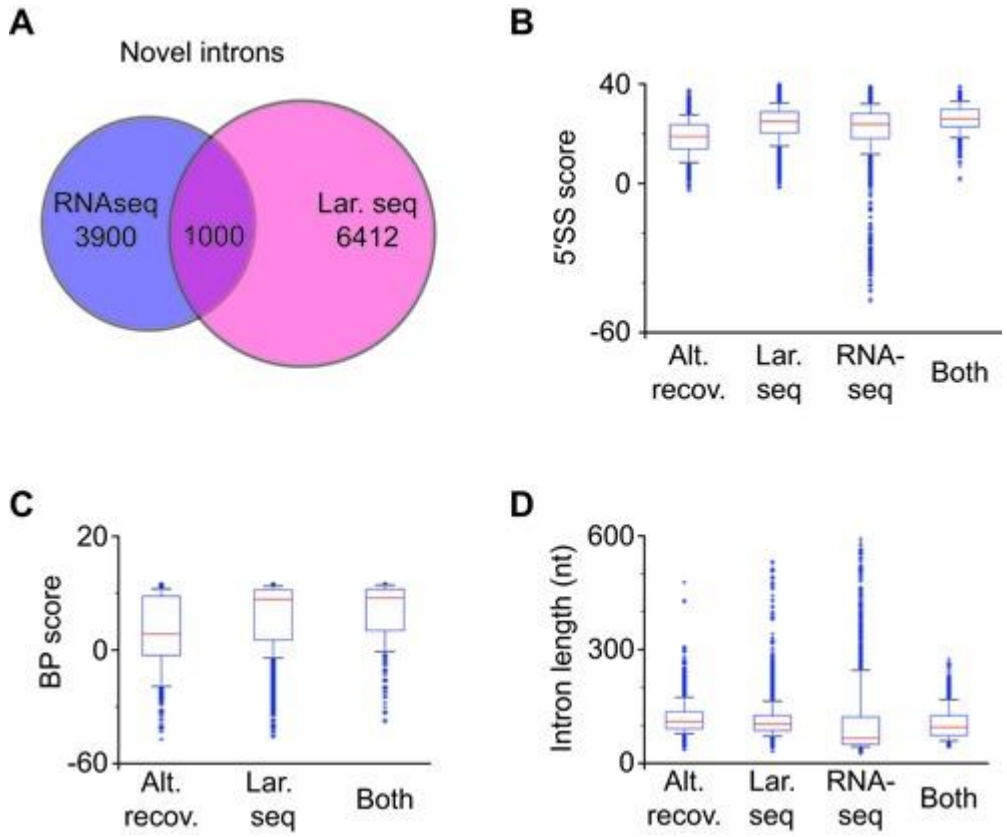


Figure A1.4 Cross validation and comparisons of novel introns detected using RNAseq and lariat sequencing. (**A**) Venn diagram illustrating novel introns identified by RNAseq, lariat sequencing, or both. (**B**) Boxplot distributions of 5′SS strength (y-axis) for 5′SSs corresponding to alternative sites recovered using lariat sequencing (alt. recov.) for annotated introns, or for novel introns identified using: lariat sequencing (lar. seq); RNAseq; or both. (**C**) Boxplot distributions of BP sequence strength (y-axis) for categories indicated in (B). (**D**) Boxplot distributions of intron length (y-axis) for categories indicated in (B).

During final preparation of our work, an analysis of publicly available RNAseq datasets from a variety of *S. pombe* growth conditions and mutants were examined for unannotated splicing events (27). As with the analysis of our own RNAseq data, we sought to compare the novel splicing events identified by lariat sequencing with those identified in this new

study. Remarkably, even though nearly 4 billion reads of RNAseq data were analyzed, comparison of these published data with our lariat sequencing generated a similar result: significant overlap of the novel splicing events was identified between the datasets, with each approach further identifying unique subsets of events. For example, of the 2923 alternative 5′SSs and 3′SSs associated with known introns identified that were identified by lariat sequencing, only ~15% were identified in the published RNAseq data (Supplementary Table S5). By contrast, 2472 events were uniquely identified by lariat sequencing and 1207 events were uniquely identified in the published RNAseq data. Importantly, as before, the majority of the events that went undetected by lariat sequencing were expected to generate lariats of sizes not readily recovered in our experiment. Similar patterns were observed when comparing exon skipping events, and novel introns (Supplementary Tables S4 and S8). Together, these results reinforce the idea that there are many additional locations within the *S. pombe*genome that are acted upon by the spliceosome but have not yet been identified by either method.

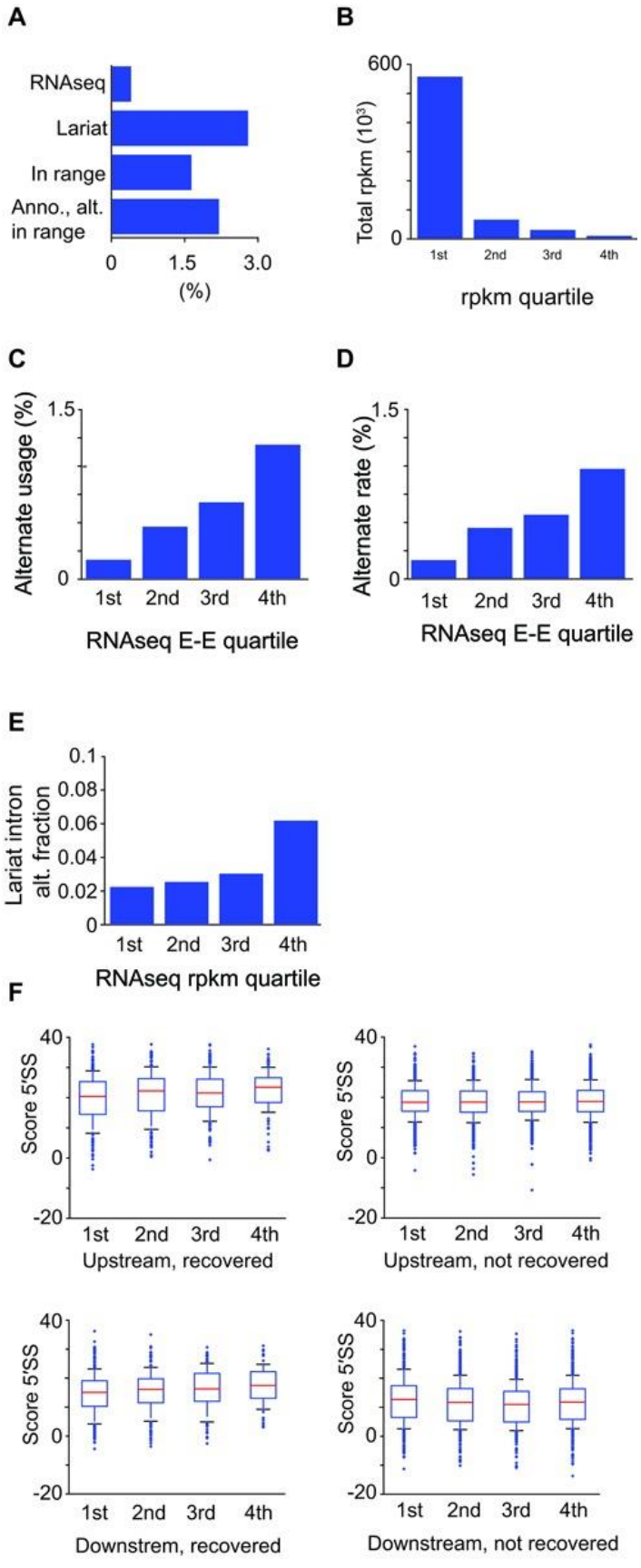
Estimating the frequency of alternative splicing

Having identified thousands of locations of alternative splicing, we next sought to characterize the frequency of these events. As a simple gauge of the extent of alternative splicing, we determined the total number of alternative splice site reads for every annotated intron relative to the sum of all annotated and alternate reads (from Supplementary Table S4). Remarkably, this yielded an alternative rate of 2.8% (Figure A1.5A), far higher than comparable values estimated from exon-exon spanning reads in most RNAseq-based studies (Figure A1.5A and references (19,27)), and moderately higher than values calculated in the

background of nuclear decay mutants (27). Importantly, when the alternative splicing percentages were recalculated considering only those alternative splicing events for which the canonical lariat was within the optimal size range for lariat sequencing, and separately only those for which both the alternative and canonical lariats were within the optimal range, estimated alternative rates were determined to be 1.6% and 2.2%, respectively, still well in excess of estimates derived from our RNAseq data (Figure A1.5A).

.

Figure A1.5 Extent of alternative splicing in *S. pombe*. (**A**) Percentage of reads corresponding to alternative splice products detected by RNAseq or lariat sequencing (lar. seq), and in lariat sequencing restricting the analysis to introns for which the annotated intron has a size optimal for detection by lariat sequencing (lar. seq opt. size; 70–150 nucleotides). (**B**) total RNAseq RPKM values (y axis) for intron containing transcripts separated into expression quartiles (x axis). (**C**) Percentage of reads corresponding to alternative splice products detected by RNAseq (y axis), shown by expression quartiles (x axis). (**D**) Inferred rate of alternative splicing, modeled using RNAseq data, shown by expression quartiles (x axis). (**E**) Percentage of reads corresponding to alternative splice products detected by lariat sequencing, shown by expression quartiles (x axis). (**F**) Boxplot distributions of 5′SS strength (y-axis) for upstream and downstream alternative 5′SSs for each expression quartile (x-axis; recovered), and for best-scoring upstream and downstream candidate sites whose usage was not observed (not recovered).


A
RNAseq
Lariat
In range
Anno., alt. in range
0
1.5
3.0
(%)
B
600
Total rpkm (10³)
0
1st
2nd
3rd
4th
rpkm quartile
C
Alternate usage (%)
1.5
0
1st
2nd
3rd
4th
RNAseq E-E quartile
D
Alternate rate (%)
1.5
0
1st
2nd
3rd
4th
RNAseq E-E quartile
E
Lariat intron alt. fraction
0.1
0.08
0.06
0.04
0.02
0
1st
2nd
3rd
4th
RNAseq rpkm quartile
F
Score 5'SS
40
0
-20
1st
2nd
3rd
4th
Upstream, recovered
Upstream, not recovered
Downstrem, recovered
Downstream, not recovered

Although the overall rate of alternative splicing detected in our RNAseq data was significantly lower than what was measured by lariat sequencing, we noted in our data that a broad range of error rates were measured among the different transcripts. In particular, alternative splicing rates were relatively low for highly expressed genes, but more pronounced for those with low expression. Therefore, to account for any relationship between gene expression and fidelity of splicing, intron containing genes were separated into expression quartiles as determined by host-transcript RPKM values (Figure A1.5B). The proportion of exon-exon spanning reads in the RNAseq data that corresponded to alternative products was then separately recalculated for each of the four expression quartiles (Figure A1.5C). In addition, a maximum-likelihood based approach was used to estimate an alternative splicing rate within each quartile of genes (Figure A1.5D). Importantly, this strategy excluded from our analysis all genes for which we detected high proportions of alternative splice products, as determined by a likelihood scoring approach, reasoning that such events are more likely to correspond to *bona fide* alternative splicing rather than errors in splicing. Together, both of these approaches showed that highly expressed genes were spliced with extremely high fidelity, whereas the fidelity of splicing decreases as expression quartiles decrease, a result that is robust to different likelihood threshold calculations (Supplementary Figure S5). Importantly, the association of decreased fidelity of splicing with more lowly expressed genes is also observed when calculated using the percentage of alternative branch-spanning reads detected in lariat sequencing (Figure A1.5E), although the extent of alternative splicing detected using lariat sequencing far exceeded that detected with RNAseq, a result consistent with lariat sequencing possessing enhanced sensitivity to detect alternative isoforms subject to rapid decay. It is worth noting that no striking differences are

apparent between the quartiles when comparing the scores of upstream or downstream alternative 5′SSs (Figure A1.5F). Moreover, the top-scoring potential alternative 5′SSs within introns for which we observed no alternative splicing were comparable to the alternative sites used (Figure A1.5F).

Given the large number of novel introns identified by both lariat sequencing and RNAseq, we determined their percent spliced index (PSI) (30), based upon our RNAseq data, in an effort to characterize whether these events reflected: *bona fide* introns whose annotations are incomplete, regulated splicing events with low PSI under standard growth conditions, or low-frequency events likely to represent splicing noise. For each novel intron, the RNAseq data were examined to identify reads spanning exon–exon or exon–intron boundaries, reflecting the spliced and unspliced isoforms, respectively. Interestingly, 93 of these introns showed a PSI of over 80%, consistent with the behavior of *bona fide*, canonical introns. These introns were distributed between coding and non-coding, sense and anti-sense transcripts, and argue for modifications of their genome annotations (Supplementary Table S8). Similarly, an additional 523 introns showed a PSI between 20% and 80%. While these PSI values were lower than expected for a canonical intron, they suggest the possibility that these are conditionally regulated splicing events. For the vast majority of the novel introns identified, PSI was below 20%. Although it is difficult to discern the functional significance of any given isoform simply on the basis of its PSI, we chose to refer to these low frequency events as ‘aberrant’.

The majority of alternative splicing events in *S. pombe* are not conserved in closely related species

To gain additional perspective on the potential functional relevance of alternative splicing in *S. pombe*, we investigated the extent to which alternative splice sites are significantly evolutionarily conserved. The PWM scores calculated for splice sites in *S. pombe* were compared with scores for the orthologous positions in three related *Schizosaccharomyces* species: *S. octosporus*, *S. cryophilus* and *S. japonicus*(Figure A1.6A) (19). As expected, annotated 5′SSs and BPs in *S. pombe* overwhelmingly maintain their splice site identity in related species (Figure A1.6B and 6C, respectively). In contrast, a comparison of the alternative sites identified by lariat sequencing in *S. pombe* showed that a large fraction (68–89%, depending on the species compared) of 5′SSs in the related species have no potential to function as splice sites (Figure A1.6Dand Supplementary Figure S6). There are, however, many sites whose sequences in related species closely match consensus 5′SSs used as alternative splice sites in *S. pombe* (Figure A1.6E and 6F, for upstream and downstream alternative 5′SSs, respectively).

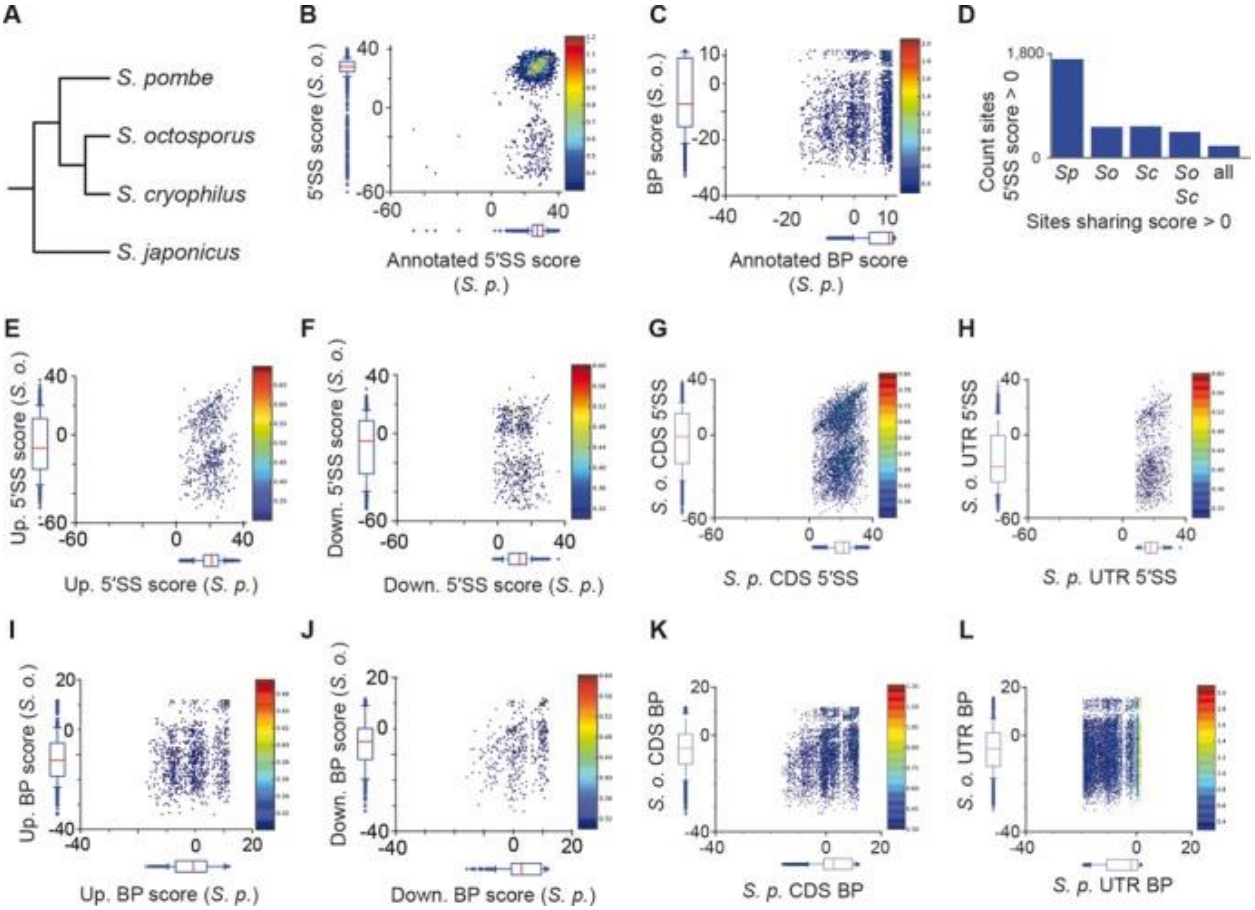


Figure A1.6 Comparative analyses of *S. pombe* splice sites. (**A**) Cladogram illustrating *Schizosaccharomyces* species included in subsequent analyses. (**B**) Heat map and boxplots indicating relationship and distributions of annotated 5′SSs in *S. pombe* (x-axis) and *S. octosporus* (y-axis). (**C**) Recovered annotated intron BP sequences, plotted as in (B). (**D**) Counts of alternative 5′SSs in *S. pombe* exceeding a score of zero, together with counts for sites conserved in indicated species. (**E**–**H**) Comparison of 5'SS between *S. pombe* and *S. octosporus* for: alternate upstream 5′SSs (**E**), alternate downstream 5'SS (**F**), control upstream 5′SSs (**G**) and control downstream 5'SSs (**H**); plotted as in (B). (**I**–**L**) Comparison of BPs between *S. pombe* and *S. octosporus* for: alternate upstream BPs (**I**), alternate downstream BPs (**J**), control upstream BPs (**K**) and control downstream BPs (**L**); plotted as in (B).

To determine the background level of conservation that exists independent of possible functions as 5′ splice sites, a similar analysis was performed on theoretical 5′SSs we found within coding sequences from genes for which no evidence of splicing exists, and separately, from noncoding sequence within 3′ untranslated regions (UTR). We reasoned that the level of conservation we detected from such theoretical 5′SSs would be a suitable background estimate for the extent of conservation we detected for real sites, and thus enable us to estimate the number of sequences selectively maintained to function as splice sites. This approach suggested that a small minority of alternate 5′SSs, both upstream and downstream, were selectively maintained above our background estimate (Figure A1.6G and H; see Supplementary Table S9 for quantification), with up to ~10% of alternate sites potentially the result of conservation, presumably corresponding to biologically meaningful occurrences of alternative splicing. In contrast, potential conservation of orthologous alternate BPs more strongly resembled the background distribution (Figure A1.6I and 6J and Supplementary Table S9). The apparent lack of conservation might be complicated by the dilution of conservation signal due to the propensity for a given 3′SS to utilize one of several possible BPs. Regardless, the lack of strong orthologous splicing signals suggested that most alternative BP usage results from aberrant splicing.

Aberrant splicing in *S. pombe*

Our discovery of widespread alternative splicing in *S. pombe*, very little of which exhibited evidence of conservation, suggested to us that the majority of the observed alternative events represented aberrant splice site usage. At a high frequency, the alternative splicing events identified by lariat sequencing generated transcripts with expected reductions

in overall stability. A total of 1661 of these alternate splicing events generated a frameshift of the resulting mRNA; an additional 376 events maintained coding frame but introduced premature stop codons in the mRNA. By contrast, only 559 of the alternative splicing events neither changed the reading frame nor introduced premature stop codons.

To better assess the stability of the aberrant splicing events we identified, additional RNAseq data were generated from a strain deficient for *upf1*, an essential component of the nonsense-mediated mRNA decay (NMD) pathway that selectively degrades erroneous transcripts (Supplementary Table S7). To determine whether alternative isoforms were stabilized in this strain, we carefully examined a subset of transcripts that satisfied three criteria: the same alternative event was identified in both datasets, the expression level of the host transcript varied by less than 25% between datasets, and the total number of alternative reads in the *Δupf1* dataset exceeded a threshold of 10 counts. As expected, when considering only those events that satisfied these criteria, the average alternate usage rate increased by over 50%, consistent with destabilization of these isoforms in wild type cells.

We considered it likely that the frequency at which an aberrant site was activated would be related to its strength as an alternative splice site. Somewhat surprisingly, however, only a weak correlation was observed between upstream alternative 5′SS scores and their usage in the wild type RNAseq dataset (Pearson correlation coefficient of 0.12, $p = 0.04$; Figure A1.7A). A weak but more significant correlation was also seen for downstream alternative 5′SSs ($\rho = 0.16$; $P = 0.01$). By contrast, in the *Δupf1* dataset the rate of utilization of both upstream and downstream alternative 5′SSs were more significantly, albeit still weakly, associated with the strength of alternative sites ($\rho = 0.22$, $P = 5 \times 10^{-4}$, and $\rho = 0.23$, $P = 5 \times$

10^{-5}, respectively). We also explored whether utilization rate of alternative sites, in RNAseq data from either wild-type or *upf1*-deficient cells, might correlate with conservation of splice sites; such analyses (Supplementary Figure S7) identified no such correlations. Taken together, these results indicate extensive utilization of a wide range of splice sites in *S. pombe*, with many of the resulting aberrant splice products substrates for cellular decay pathways.

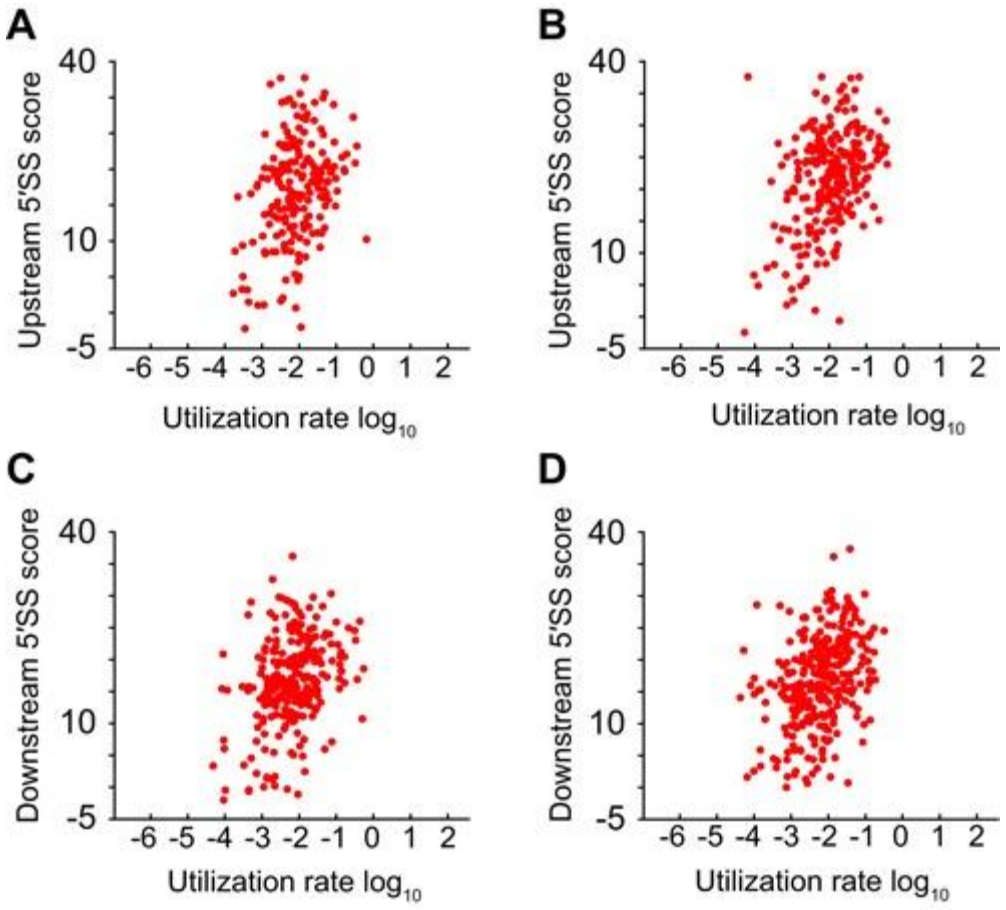


Figure A1.7 Influence of splice site strength on frequency of alternative splicing. (**A–D**) Scatter-plots of proportion of alternative splicing events found in RNAseq data (x-axis) plotted against predicted strength of alternative 5′SS (y-axis). Alternative events analyzed separately for upstream (**A**, **B**) and downstream alterative 5′SSs (**C**, **D**), using RNAseq from wild-type (**A**, **C**) and NMD-deficient strains (**B**, **D**).

DISCUSSION

Pre-mRNA splicing is a critical component of eukaryotic gene expression. By temporally regulating the activation of different splice sites within a transcript, the process of alternative splicing provides a powerful opportunity for organisms to expand their proteomic diversity (4). The importance of appropriate splice site selection is highlighted by the number of human diseases that are associated with mutations at or near these sites (31,32). Attempts to understand the rules that govern splice site selection have largely been driven by analysis of mRNAs present in cells, and inferences about the splice sites that were used to generate those mRNAs (33); however, a major shortcoming of this approach is its failure to detect those splicing events that lead to destabilization of the spliced product.

Here, we have used lariat sequencing to enable an understanding of splicing in *S. pombe* not possible with RNAseq. As a tool for elucidating the diversity of substrates acted upon by the spliceosome, the major advantage of lariat sequencing derives from directly sequencing introns, rather than sequencing mature transcripts that are subject to RNA quality control pathways. Because lariat sequencing directly identifies both the 5′SS and the BP used during the splicing reaction, the data generated here represent the largest collection of experimentally identified splice sites in an organism with degenerate splice sites, allowing an opportunity to visualize the sequence constraints of the spliceosome in a way not previously possible.

Our data reveal a remarkable level of alternative and aberrant splice site activation across the *S. pombe*genome, including nearly 3000 examples of alternative splice site activation surrounding annotated introns, hundreds of novel examples of exon skipping, and thousands

of examples of novel introns (Figure A1.2). Importantly, because of the length limitations of lariat sequencing, these events underrepresent the total number of alternative and aberrant splicing events which must exist in the *S. pombe* genome. Together, our data suggest that rates of alternative splice site activation in *S. pombe* are around 3%, a value significantly higher than previous estimates (9,10,19). The difference between our lariat derived error estimates and published RNAseq estimates is particularly noteworthy given the expectation that alternative splicing in the less complex genome of *S. pombe* is predicted to be lower than that seen in mammals (34).

The instances of alternative 5′ and 3′SS activation identified here are remarkable not only because of the frequency with which they occur, but also because of the relative strengths of the sites that are being activated. Indeed, given the strength of many of these alternative site sequences, the more surprising finding may be that they aren't activated at higher frequency. This observation, along with the failure to identify alternative splicing at many high scoring cryptic sites, underscores the significance of context in understanding splice site strength. The identification here of both activated and silent cryptic splice sites offers a powerful opportunity to better understand the constraints that drive splice site activation. Future experiments examining the subsets of activated and silent sites should provide insights into the mechanisms by which cryptic splice sites can be activated or repressed.

Our analysis of exon skipping events also provides a surprising insight into the mechanism of spliceosome assembly in *S. pombe*. In higher eukaryotes, where introns can be exceptionally large and exons tend to be shorter, strong evidence exists in support of an ‘exon-definition’ model for spliceosome assembly, wherein recognition of the 5′SSs by the

U1 snRNP can facilitate recognition of an upstream BP/3′SSs (35). By contrast, in lower eukaryotes like *S. pombe*, where introns are much shorter, assembly is thought to occur across introns, agnostic to the content of surrounding introns (28,36). In this context, our finding that exon skipping events tended to be enriched for poor scoring 5′SSs in the downstream introns was completely unexpected (Figure A1.2D). Interestingly, the simple model that the subset of introns identified here represent the few splicing events in *S. pombe* that utilize exon-definition for spliceosome assembly seems unlikely since the skipped BP sequences have scores which are largely indistinguishable from the global population. Alternatively, these results suggest the possibility that cross-exon interactions facilitate spliceosome assembly for many or all introns, including those presumed to assemble primarily by intron-definition, and that exons with weak downstream 5′SSs are more likely to exhibit exon skipping because of the decreased ability to utilize these cross-exon interactions.

Our results demonstrate that estimates of alternative splicing using RNAseq alone are likely to significantly underestimate the prevalence of alternative splicing. Splicing errors that generate aberrant transcripts will be significantly underestimated by RNAseq because they are likely to be subjected to RNA degradation pathways, including nonsense-mediated mRNA decay (NMD) and spliceosome-mediated decay (SMD) (37,38). Recent work in budding yeast, where splice sites conform to a strong consensus sequence, also revealed an unappreciated level of alternative splice site selection, much of which is masked by the NMD pathway (39). Similarly, recent work in *S. pombe* identified widespread alternative splicing at rates approaching those detected here in the background of nuclear decay mutants (27).

Moreover, although we have associated the lariats identified here with splicing events that have completed both chemical steps, a fascinating example of biologically-relevant, first-step only splicing has been demonstrated for the TER1 transcript in *S. pombe* (40). As such, we cannot preclude the possibility that some of the lariats identified here are the products of reactions that have only undergone the first chemical step of the splicing pathway. Additional experiments will be necessary to fully understand the mechanisms by which these alternative splicing events are generated and subsequently linked with cellular decay pathways.

A surprising consequence of sequencing intron lariats was our discovery that lariats derived from different introns of multi-intronic genes have highly discrepant abundances, both as measured by lariat sequencing and confirmed with qRT-PCR and RNAseq. In organisms from yeast to humans, and including *S. pombe*, the nuclear processing of many non-coding RNAs is accomplished through endonucleolytic cleavage by RNase III homologs (Pac1 in *S. pombe*) (41). Recent work in budding yeast demonstrates that Rnt1, the homolog of Pac1, cleaves more targets than previously expected (38,42); it remains unknown whether Pac1 or an as yet unidentified endonuclease contributes to the degradation of lariat introns.

While our data make it clear that alternative splicing is widespread in *S. pombe*, from the perspective of *S. pombe* biology, it is less clear that these events are functionally significant. Sites that have been selectively maintained over evolutionary time likely correspond to biologically meaningful alternative splicing events, whereas sites that have diverged at a neutral evolutionary rate are more likely to correspond to errors in splicing (43,44). Our comparative analyses of alternative splice site sequences indicate that the preponderance of alternative splicing in *S. pombe* has not been maintained, even in closely related species. In

the absence of evolutionary conservation, we conclude that the majority of the alternative splice sites we detected in *S. pombe* correspond either to rapidly evolving functional splicing events in *S. pombe*, likely true for only a very small subset of sites, or splicing errors that have arisen as a consequence of neutral genome evolution in the *S. pombe* lineage. Nevertheless, although the majority of the alternative events detected here likely have no physiological function in *S. pombe*, the widespread aberrant splicing identified here almost certainly plays an important role in genome evolution. Presumably, permissive alternative splicing, typically resulting in aberrant transcripts that are selectively degraded, provides the raw material from which advantageous events are selected during evolution. Our study suggests that the error rate intrinsic to splicing, acting upon cryptic splice sites, greatly exceeds previous estimates, perhaps facilitating more rapid acquisition of conserved alternative splicing events.

Although lariat sequencing is not readily amenable to sequencing introns from higher eukaryotes, chiefly because electrophoretic separation of large lariats is impractical, it will be important to develop variations on this approach that are suitable for intron sequencing from any species. Given the conservation of the splicing apparatus, and the overall similarity of splice site sequences between *S. pombe* and humans, we predict that a similar, or higher level of aberrant splice site activation will occur in humans as well. Knowledge of the locations and identities of splice sites activated in the human genome, together with the information derived here, will help in better understanding the mechanistic bases of splice site selection.

ACKNOWLEDGMENTS

We would like to acknowledge members of the Pleiss and Grimson labs for critical input throughout the project. This work was funded by NIGMS grants GM098634 to JAP and funds provided by Cornell University to AG.

Authors Contributions: Experiments were conceived and designed by N.S., M.R., E.A.F., A.G., and J.A.P. Experiments were performed by N.S., M.R., and E.A.F. Data were analyzed by N.S., M.R., A.G., and J.A.P. The manuscript was written by N.S., A.G. and J.A.P.

FUNDING

National Institutes of Health (NIH) [GM098634]. Funding for open access charge: NIH [GM098634].

Conflict of interest statement. - None declared.

SUPPLEMENTARY DATA

All Supplementary data are available at NAR Online.

REFERNCES

1. Lee Y. Rio D.C. Mechanisms and regulation of alternative pre-mRNA splicing Annu. Rev. Biochem. 2015 84 291 323

2. Moore M. Query C. Sharp P. Splicing of precursors to messenger RNAs by the spliceosome The RNA World 1993 1 NY Cold Spring Harbor Laboratory Press 303 357

3. Domdey H. Apostol B. Lin R.-J. Newman A. Brody E. Abelson J. Lariat structures are in vivo intermediates in yeast pre-mRNA splicing Cell 1984 39 611 621

4. Nilsen T.W. Graveley B.R. Expansion of the eukaryotic proteome by alternative splicing Nature 2010 463 457 463

5. Chen M. Manley J.L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches Nat. Rev. Mol. Cell Biol. 2009 10 741 754

6. Lim L.P. Burge C.B. A computational analysis of sequence features involved in recognition of short introns Proc. Natl. Acad. Sci. U.S.A. 2001 98 11193 11198

7. Hamid F.M. Makeyev E.V. Emerging functions of alternative splicing coupled with nonsense-mediated decay Biochem. Soc. Trans. 2014 42 1168 1173

8. Fox-Walsh K.L. Hertel K.J. Splice-site pairing is an intrinsically high fidelity process Proc. Natl. Acad. Sci. U.S.A. 2009 106 1766 1771

9. Melamud E. Moult J. Stochastic noise in splicing machinery Nucleic Acids Res. 2009 37 4873 4886

10. Pickrell J.K. Pai A.A. Gilad Y. Pritchard J.K. Noisy splicing drives mRNA isoform diversity in human cells PLoS Genet. 2010 6 e1001236

11. Wood V. Gwilliam R. Rajandream M.-A. Lyne M. Lyne R. Stewart A. Sgouros J. Peat N. Hayles J. Baker S. et al. The genome sequence of Schizosaccharomyces pombe Nature 2002 415 871 880

12. Kuhn A.N. Käufer N.F. Pre-mRNA splicing in Schizosaccharomyces pombe Curr. Genet. 2002 42 241 251

13. Awan A.R. Manfredo A. Pleiss J.A. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans Proc. Natl. Acad. Sci. U.S.A. 2013 110 12762 12767

14. Gilbert W. Why genes in pieces? Nature 1978 271 501 501

15. Forsburg S.L. Rhind N. Basic methods for fission yeast Yeast Chichester Engl. 2006 23 173 183

16. Bolger A.M. Lohse M. Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data Bioinforma. Oxf. Engl. 2014 30 2114 2120

17. Langmead B. Salzberg S.L. Fast gapped-read alignment with Bowtie 2 Nat. Methods 2012 9 357 359

18. Taggart A.J. DeSimone A.M. Shih J.S. Filloux M.E. Fairbrother W.G. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo Nat. Struct. Mol. Biol. 2012 19 719 721

19. Rhind N. Chen Z. Yassour M. Thompson D.A. Haas B.J. Habib N. Wapinski I. Roy S. Lin M.F. Heiman D.I. et al. Comparative functional genomics of the fission yeasts Science 2011 332 930 936

20. Crooks G.E. Hon G. Chandonia J.-M. Brenner S.E. WebLogo: a sequence logo generator Genome Res. 2004 14 1188 1190

21. Bitton D.A. Rallis C. Jeffares D.C. Smith G.C. Chen Y.Y.C. Codlin S. Marguerat S. Bähler J. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq Genome Res. 2014 24 1169 1179

22. Mercer T.R. Clark M.B. Andersen S.B. Brunck M.E. Haerty W. Crawford J. Taft R.J. Nielsen L.K. Dinger M.E. Mattick J.S. Genome-wide discovery of human splicing branchpoints Genome Res. 2015 25 290 303

23. Wood V. Harris M.A. McDowall M.D. Rutherford K. Vaughan B.W. Staines D.M. Aslett M. Lock A. Bähler J. Kersey P.J. et al. PomBase: a comprehensive online resource for fission yeast Nucleic Acids Res. 2012 40 D695 D699

24. Aebi M. Weissman C. Precision and orderliness in splicing Trends Genet. 1987 3 102 107

25. Webb C.J. Romfo C.M. van Heeckeren W.J. Wise J.A. Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin Genes Dev. 2005 19 242 254

26. Pérez-Valle J. Vilardell J. Intronic features that determine the selection of the 3′ splice site Wiley Interdiscip. Rev. RNA 2012 3 707 717

27. Bitton D.A. Atkinson S.R. Rallis C. Smith G.C. Ellis D.A. Chen Y.Y.C. Malecki M. Codlin S. Lemay J.-F. Cotobal C. et al. Widespread exon skipping triggers degradation by nuclear RNA surveillance in fission yeast Genome Res. 2015 doi:10.1101/gr.185371.114

28. Romfo C.M. Alvarez C.J. van Heeckeren W.J. Webb C.J. Wise J.A. Evidence for Splice Site Pairing via Intron Definition in Schizosaccharomyces pombe Mol. Cell. Biol. 2000 20 7955 7970

29. Kim D. Pertea G. Trapnell C. Pimentel H. Kelley R. Salzberg S.L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions Genome Biol. 2013 14 R36

30. Venables J.P. Klinck R. Bramard A. Inkel L. Dufresne-Martin G. Koh C. Gervais-Bird J. Lapointe E. Froehlich U. Durand M. et al. Identification of Alternative Splicing Markers for Breast Cancer Cancer Res. 2008 68 9525 9531

31. Cooper T.A. Wan L. Dreyfuss G. RNA and disease Cell 2009 136 777 793

32. Xiong H.Y. Alipanahi B. Lee L.J. Bretschneider H. Merico D. Yuen R.K.C. Hua Y. Gueroussov S. Najafabadi H.S. Hughes T.R. et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease Science 2015 347 1254806

33. Wang Z. Burge C.B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code RNA N. Y. N 2008 14 802 813

34. Kim E. Magen A. Ast G. Different levels of alternative splicing among eukaryotes Nucleic Acids Res. 2007 35 125 131

35. De Conti L. Baralle M. Buratti E. Exon and intron definition in pre-mRNA splicing Wiley Interdiscip. Rev. RNA 2013 4 49 60

36. Shao W. Kim H.-S. Cao Y. Xu Y.-Z. Query C.C. A U1-U2 snRNP interaction network during intron definition Mol. Cell. Biol. 2012 32 470 478

37. Volanakis A. Passoni M. Hector R.D. Shah S. Kilchert C. Granneman S. Vasiljeva L. Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast Genes Dev. 2013 27 2025 2038

38. Roy K. Chanfreau G. Stress-induced nuclear RNA degradation pathways regulate yeast bromodomain factor 2 to promote cell survival PLoS Genet. 2014 10 e1004661

39. Kawashima T. Douglass S. Gabunilas J. Pellegrini M. Chanfreau G.F. Widespread Use of Non-productive Alternative Splice Sites in Saccharomyces cerevisiae PLoS Genet 2014 10 e1004249

40. Box J.A. Bunch J.T. Tang W. Baumann P. Spliceosomal cleavage generates the 3′ end of telomerase RNA Nature 2008 456 910 914

41. Chanfreau G. Conservation of RNase III Processing Pathways and Specificity in Hemiascomycetes Eukaryot. Cell 2003 2 901 909

42. Gagnon J. Lavoie M. Catala M. Malenfant F. Elela S.A. Transcriptome wide annotation of eukaryotic RNase III reactivity and degradation signals PLoS Genet 2015 11 e1005000

43. Merkin J. Russell C. Chen P. Burge C.B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues Science 2012 338 1593 1599

44. Barbosa-Morais N.L. Irimia M. Pan Q. Xiong H.Y. Gueroussov S. Lee L.J. Slobodeniuc V. Kutter C. Watt S. Colak R. et al. The evolutionary landscape of alternative splicing in vertebrate species Science 2012 338 1587 1593

APPENDIX 2

BRAIN REGION-SPECIFIC TEMPORAL EXPRESSION PROFILES OF PLASTICITY-RELATED PROTEIN TRANSCRIPTS INDUCED BY OLFACTORY ASSOCIATIVE LEARNING

Michelle T. Tong[1], Madhura Raghavan[2], Jeffrey A. Pleiss[2], and Thomas A. Cleland[1]

[1]Department of Psychology

[2]Department of Molecular Biology and Genetics

Cornell University, Ithaca NY 14853, USA

This work was a collaboration with Michelle Tong and Thomas Cleland in the Department of Psychology at Cornell. The idea behind the project was conceived by Michelle Tong and Thomas Cleland as a part of their study on the temporal and spatial profiles of the molecular mechanisms involved in long-term memory. A collaboration was forged with our lab to help with profiling the transcriptional changes in brain samples collected at different points in the formation of long-term memory. I was involved in designing, performing and analyzing the data for transcriptional profiling of select targets using RT-qPCR. The chapter presented here describes the overall study with the portion of the work (RT-qPCR) I was involved in complete detail.

ABSTRACT

Previous research has contributed to a strong understanding of the role of many molecular mechanisms involved in long-term memory (LTM), including describing the effects of those mechanisms on longer-term structural changes. This research has also suggested that LTM depends as much on the temporal specificity of these mechanisms as on their downstream effects. It is yet unclear how the timing of these mechanisms as well as their coordinated activity across the multiple brain regions is involved in learning. In the present study, we took a step towards characterizing the time course of several molecular mechanisms across multiple brain regions. We trained mice on an associative odor learning task for 1, 2, 4, or 6 days. We collected the OB, striatum, hippocampus, cortex, and cerebellum from the mice on each day prior to training, immediately after training, or 15, 30, or 60 minutes after training. We then used high-throughput RT-PCR to analyze mRNA levels for several plasticity-related proteins (PRPs), including *bdnf*, intracellular signaling cascades, *erk1* and *erk2*, transcription factor, *creb1*, and immediate early genes, *arc*, *fos*, and *erg1*. We found that learning-responsive transcription differed between genes, and that PRP time courses differed as a function of brain region.

INTRODUCTION

Molecular mechanisms (including intracellular cascades, molecular signaling, neuromodulatory influences, activity-dependent protein synthesis, and epigenetic modifications) have been shown to underlie LTM consolidation and persistence through their role in long-term structural changes (such as long-term potentiation or other synaptic weight changes, alterations to neuronal morphology such as dendritic branching, changes to terminal shapes or numbers, and even ancillary modifications such as effects on glia or cell adhesion to the extracellular matrix, and changes to neuron number via adult neurogenesis or selective apoptosis).

A great deal of work, particularly those using the inhibitory avoidance (IA) paradigm, has contributed to our knowledge of the functional roles of these molecular mechanisms. IA conditioning leads to a rapid elevation of several intracellular signaling pathways in the hippocampus (HPC), including CaMKII, CAMP, and the LTM-associated transcription factor CREB (Ferrer et al., 1996). Learning promotes the trafficking of receptors in the plasma membrane, including ionotrophic glutamate receptors (Danysz et al., 1995; Riedel et al., 1995), and increases the transcription of other plasticity- related mechanisms, like BDNF (Chaudhury et al., 2010).

Of particular relevance to our present discussion is the fact that work examining the activity of these mechanisms suggests that LTM consolidation relies as much on the temporal specificity of the mechanisms as it does on their downstream effects. For example, blocking CaMKII activity immediately after IA training substantially reduced animals' fear

responses when measured 24 hours later (i.e., LTM). However, blocking CaMKII activity 30 minutes after IA resulted in a weaker LTM deficit, and blockade 2-4 hours after IA had no effect on LTM at all (Wolfman et al., 1994). Studies that directly measured CaMKII levels revealed that CaMKII activity increased immediately after IA training, and remained high when tested 30 minutes later, but had returned to baseline when tested two hours after conditioning (Bernabeu et al., 1997a). These findings indicate that CaMKII plays a crucial role early in the memory induction process, and that its functional role in LTM formation is confined to a specific period following learning. Similar results for other mechanisms, including PKA (Bernabeu et al., 1996, Bernabeu et al., 1997b, Bernabeu et al., 1997a, Bevilaqua et al., 1997, Cammarota et al., 1997, Izquierdo et al., 1997, Jerusalinsky et al., 1994, Wolfman et al., 1994), NMDA and AMPA (Bernabeu et al., 1997a, Izquierdo et al., 1992, Jerusalinsky et al., 1992), and BDNF (Bekinschtein et al., 2007), have been previously described. .

In addition to the importance of understanding timecourse, limited studies have also suggested that timecourses for the mechanisms differ across brain regions in meaningful ways. For example, Izquierdo and colleagues (1997) gave infusions of either NDMA or AMPA receptor antagonists into the amygdala, HPC, and entorhinal cortex 0, 90, 180, and 360 minutes after IA training. Amongst other results, the authors found that 24-hour memory was significantly attenuated in animals that received amygdala and HPC infusions of the NMDA receptor antagonist, AP5, immediately after training. Twenty-four hour memory was unimpaired in animals that received this infusion into the entorhinal cortex. However, 90- or 180-minute infusions into the entorhinal cortex did have amnesiac effects. The finding

implies that NMDA receptor activity in these areas have functionally different time courses.

Studies of other plasticity-related mechanisms in multi-trial, appetitive learning paradigms, such as a radial arm maze task, have shown that time-courses can differ with one-trial learning. In the HPC, BDNF levels were found to increase immediately after IA training. Infusions of function-blocking BDNF antisense oligonucleotides into the HPC just prior to training blocked LTM consolidation, while sparing STM, indicating that BDNF activity immediately following learning sets the stage for eventual LTM consolidation (Alonso et al., 2005, Bekinschtein et al., 2007). Similarly, BDNF mRNA levels in the HPC increased significantly after conditioning on an appetitive radial arm maze task – but only on the eighth day of training (Mizuno et al., 2002). Additionally, while BDNF mRNA levels increased 15 minutes after the last training trial in the HPC, no increase was observed during this time in the frontal cortex (Mizuno et al., 2000). These findings suggest that, first, similar molecular mechanisms are involved in multi-trial appetitive learning as in single-trial, fear-based learning, but, crucially, the timecourse after learning differs dramatically (that is, on the order of days, as opposed to minutes or hours). Second, they corroborate results that, indeed, the timecourse of plasticity-related activity differs between different brain regions.

In addition to relying primarily on one-trial learning, a weakness of previous studies is that very few have examined the mechanisms across regions for the same learning event. In the present study, we take a first step toward characterizing the timecourse of several molecular mechanisms across multiple brain regions during the same multi-trial learning event. We trained mice on an associative odor learning task for 1, 2, 4, or 6 days. We collected the OB, striatum, hippocampus, prefrontal cortex, and cerebellum from the mice on

each day prior to training, immediately after training, or 15, 30, or 60 minutes after training. We then used high-throughput RT-PCR to analyze mRNA levels for several plasticity-related proteins (PRPs), including *bdnf* (and intracellular signaling cascades, *erk1* and *erk2*), *arc*, *fos*, *erg1*, and *creb1* in these tissues.

We found that learning-responsive transcription differed between genes, and that PRP timecourses differed between brain regions. Future studies could use this "spatiotemporal" map to discover the functional consequences of these patterns.

MATERIALS AND METHODS

Animals

A total of 60 adult male CD-1 mice (Charles River), 8 weeks old at the beginning of the shaping period, were used in this study. All procedures were performed under the auspices of a protocol approved by the Cornell University Institutional Animal Care and Use Committee (IACUC). Cornell University is accredited by The Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC).

The mice were kept on a 12:12 hour reverse light/dark cycle and allowed free access to water at all times. They were kept on a food-restriction schedule designed to keep them around 90% of their free feeding weight for the duration of the behavioral experiments. This food restriction schedule began 3 days prior to the beginning of the shaping period.

Behavioral conditioning

Apparatus. The mice were tested in a clear Plexiglas cage (28 cm long x 17 cm wide x 12 cm high), bisected by a removable opaque black center divider into a *resting chamber* and a *test chamber*. Mice were placed into the resting chamber at the beginning of each session. Each trial began with the lifting of the divider, after which animals entered the test chamber and encountered the odor stimuli. Mice were returned to the resting chamber at the end of each trial.

Odor sets. For behavioral shaping, (+/-)-limonene (Sigma-Aldrich, St. Louis, MO, USA) was used as the rewarded odor and unscented mineral oil as the unrewarded odor. For training during the experiment proper, butanoic acid and pentanoic acid (two moderately

similar odorants) were used as the rewarded and unrewarded odors; the rewarded odor was counterbalanced among the mice in order to neutralize the effects of any intrinsic odor preferences. All odorants were diluted in mineral oil so as to emit a theoretical steady-state vapor phase partial pressure of 0.1 Pa (Cleland et al., 2002) and mixed into play sand at a ratio of 400 ul diluted odorant per 100 g sand. This relatively low odor concentration was chosen to limit the learning rate so that animals would not reach asymptotic performance after the first day of training.

Shaping. The mice underwent a ten-day behavioral shaping period prior to the start of the experiment proper. Each mouse was brought into the procedure room and handled for 10 minutes per day for two days (Days 1 and 2). On Day 3, a petri dish (Pyrex, 60 mm diameter, 15 mm height) filled with ~25 g of ±-limonene-scented play sand (Quikrete; Atlanta, GA) and 10-15 5-mg sucrose reward pellets (PJ Noyes Precision Pellets; TestDiet, Richmond, IN) was placed into their home cages. The scented sand and pellets were replenished on Day 4. On Days 5-6, the mice were acclimated to the behavioral apparatus. Two dishes of sand (one scented with ±-limonene and one with plain mineral oil), were placed into the Plexiglas cage without the center divider; ten reward pellets were mixed into the limonene-scented dish. Each mouse was placed into the test chamber for 10 minutes, and was allowed to explore freely and to consume the reward pellets.

On Day 7, the mice were introduced to a shortened version of the final testing procedure. The center divider was restored to the test cage, and two dishes of sand were placed into the test chamber. A single reward pellet was place on top of the limonene-scented sand. The mouse was placed into the resting chamber, after which the center divider was lifted and the mouse

was allowed to enter the test chamber and retrieve the reward pellet. Animals were returned to the resting chamber either after they retrieved the reward pellet or after 5 minutes elapsed. This was repeated for 10 trials. On any given trial, the limonene-scented, rewarded dish was randomly placed on the left or right according to a random number generator. This procedure was repeated on Day 8, except that on Day 8 the reward pellet was buried progressively more deeply in the sand with each trial. All mice were digging for an unseen reward pellet by the tenth trial on Day 8.

On Day 9, animals were presented with the full 20-trial version of the task. As on Day 8, reward pellets were fully buried under the sand in the dish; however, each trial lasted only 1 minute. Animals were allowed to dig freely in both dishes for the reward. On Day 10, the animals underwent the same 20 trials, but they were not permitted to self-correct; if a mouse dug in the unrewarded dish (plain mineral oil) first, they were returned to the resting chamber without reward and the next trial began. Mice that dug first in the limonene-scented sand were allowed to retrieve the reward pellet before being returned to the resting chamber.

Generation of odor discrimination learning curves. After shaping, for the experiment proper, the mice were divided into four training groups, which received either one, two, four, or six days of training, respectively (Figure A2.1). At the beginning of each day of training, the animals were placed into the resting chamber. Two dishes of sand, scented with butanoic acid and pentanoic acid respectively, were placed in the test chamber. A reward pellet was buried in one of the dishes. If animals dug in the rewarded odor first, they were allowed to retrieve the sugar reward and then returned to the resting chamber, and the next trial began immediately. If the animal dug in the unrewarded odor first, they were returned to the resting

chamber without reward and the next trial began immediately. One day of training comprised 20 of these trials in direct succession. Because the two odors were novel, initial performance on Day 1 was at chance, but improved over the course of the 20 trials within a given day and over the course of multiple days of training. Mice were trained for up to six consecutive days (Figure A2.2), though tissue samples were only taken on days 1, 2, 4, and 6. The maximum trial duration was one minute, after which the mouse was returned to the resting chamber and the next trial initiated.

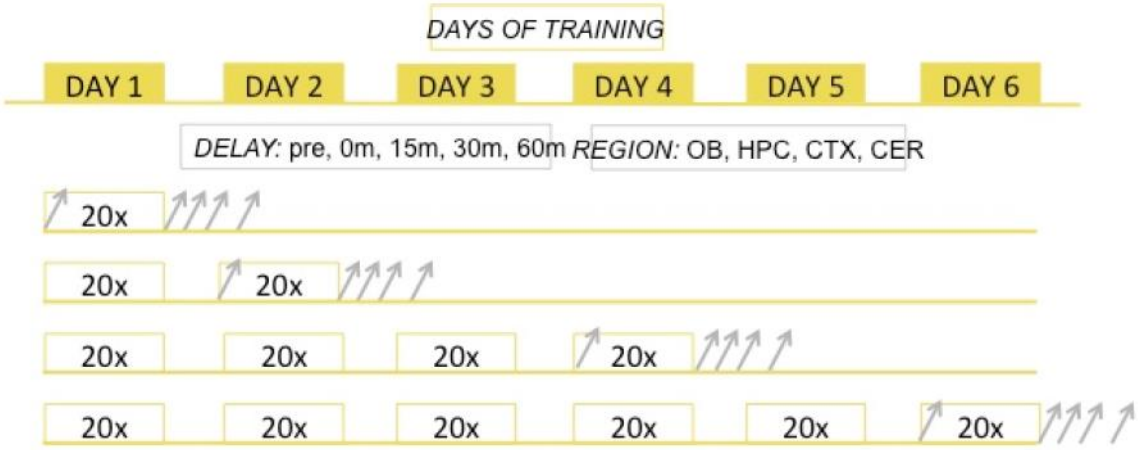


Figure A2.1: Associative learning protocol used to obtain mRNA data. Animals were given 20 trials ("20x") of the odor-reward task across 1, 2, 4, or 6 days. Grey arrows represent the "DELAY"s when brain tissue was harvested (immediately prior to behavioral training, immediately after, 15 minutes after, 30 minutes after, or 60 minutes after their final training trial. At those delays, the "REGION"s removed were the olfactory bulb (OB), the striatum (STR), the hippocampus (HPC), the cerebellum (CER), and roughly the prefrontal cortex (CTX).

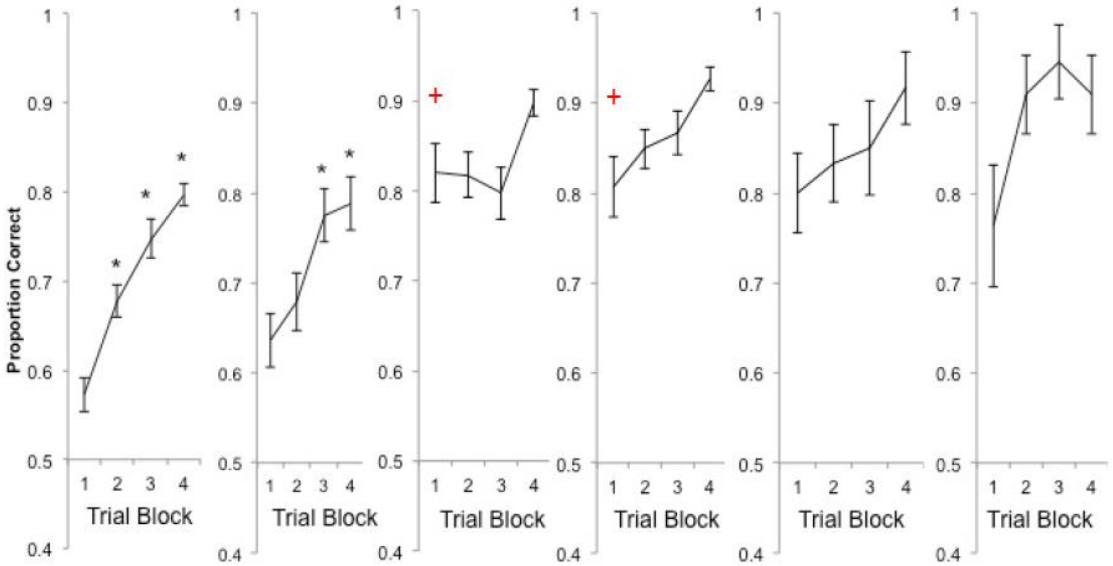


Figure A2.2: Animals show greatest learning during the first and second day of training. Stars represent Trial Blocks (TBs) which are significantly higher than the first TB of each day ($p < .05$). Performance asymptotes on the third day of training.

Tissue collection

Animals were euthanized by cervical dislocation and rapid decapitation at one of several time points on a given training day: (1) immediately prior to behavioral training, (2) immediately after the end of training, or (3) 15 minutes, (4) 30 minutes, or (5) 60 minutes after the end of training (Figure A2.1). For the 15-, 30-, and 60-minute latencies, the mice were placed back into their home cages prior to euthanasia. After decapitation, the brains

were immediately removed and bilateral dissections were performed to isolate the olfactory bulbs (OB), the striatum (STR), the hippocampus (HPC), the cerebellum (CER), and a substantial piece of isocortex (CTX). Brain tissue samples were flash-frozen on dry ice.

This process was repeated on training days 1, 2, 4, and 6 (Figure A2.1), resulting in a total of 20 timepoints at which brain tissues were sampled. 1-3 biological replicates were generated per time point per day, such that the total number of tissue samples was 300, drawn from 60 different mice.

mRNA quantification

RNA extraction and purification. Samples (quantity) of the flash-frozen brain tissue were homogenized in 2 ml Trizol. A quarter (500ul) of the homogenized tissue samples in Trizol were taken and laid out into four 96-well plates and frozen at -20^{o} C. The samples were grouped in such a way that two biological replicates of all time points for a particular tissue were present in a single plate to enable within tissue comparisons. Additional biological replicates were grouped together on another plate to enable comparison across tissues for each of the sampling timepoints. We added 6 control samples to all the plates in order to control for any technical variation arising from the handling of the plates. On the day of RNA extraction, the plates were thawed at room temperature for an hour. 100 µl of chloroform was added to each sample, vortexed and set on bench top for 5 minutes until two phases became visible in Trizol. The plates were then spun at 4000 g at 4^{o} C for 20 minutes. Using robotics, around 180 µl of the aqueous phase was transferred into a new 96-well plate. 1 ml of binding buffer (2M guanidine hydrochloride, 75% isopropanol) was added to the aqueous phase and RNA was extracted using a 96- well glass fiber binding plate (Nunc

278010) with two successive washes with 800 µl wash buffer (80% ethanol, 10mM Tris pH 8) followed by a dry spin. The RNA was then eluted twice successively in 50 µl of water per elution. The RNA yield was highest for cortical samples (~13 µg) and lowest for striatal samples (~1 µg).

cDNA synthesis. We used 80 µl of the 100 µl of purified RNA for cDNA synthesis. To 80 µl of RNA, 40 µl of primer mix (150 mM Tris-HCl pH 8.0, 225 mM KCl, 1 µg/µl dN_9 random 9-mers) was added. The samples were incubated at 65°C for 5 minutes followed by incubation on ice for 5 minutes. Then, 40 µl of of RT-mix (50 mM Tris-HCl (pH 8.0), 75 mM KCl, 12 mM $MgCl_2$, 40 mM DTT, 2 mM each dNTP, and M-MLV RT) was added. Reactions were incubated overnight at 42°C. On average, the cDNA yield was around 25%, with the CTX samples yielding ~3 µg cDNA and STR samples yielding ~300 ng cDNA. The cDNA samples were then diluted 2.5 fold to 400 µl with an effective concentration between ~0.75 ng/µl (STR) and ~7.5 ng/µl (CTX).

qPCR and analysis. The qPCR reactions were performed in a reaction volume of 10 µl, containing 5 µl of cDNA, 10 mM Tris-HCl (pH 8.5), 50 mM KCl, 1.5 mM MgCl2, 0.2 mM each dNTP, 0.25x SYBR Green, 5% DMSO, Taq DNA polymerase, and 250 nM forward and reverse primers. The primer sequences for each targeted gene is shown in Table A2.1. Primers were designed against a constitutive exon in each gene that is present in most of the transcript isoforms that were annotated on the UCSC browser. Standard curves were generated using eight 4-fold serial dilutions of cortex cDNA covering a range from 200 ng to 12 pg of cDNA (Figure A2.3). Each primer pair was well-behaved, showing amplification efficiency between 92% and 99%.

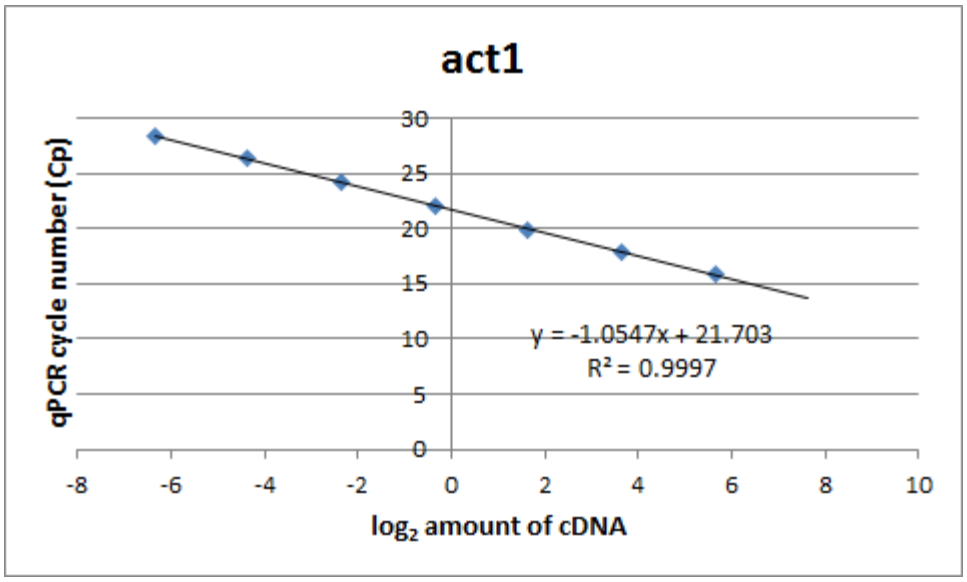
For the qPCR experiment, three technical replicates were measured for each biologically independent sample, generating up to nine independent measurements for each sample. For each technical measurement, an amount value (arbitrary unit) of each tissue was calculated using the standard curves and averages calculated for each gene. We used an existing R package to determine which of the 3 housekeeping genes (β-actin, GAPDH, or PGK1) we used was best for normalization. In these analyses, we used β-actin to normalize each sample. Analyses were then done using this normalized "amount" measure.

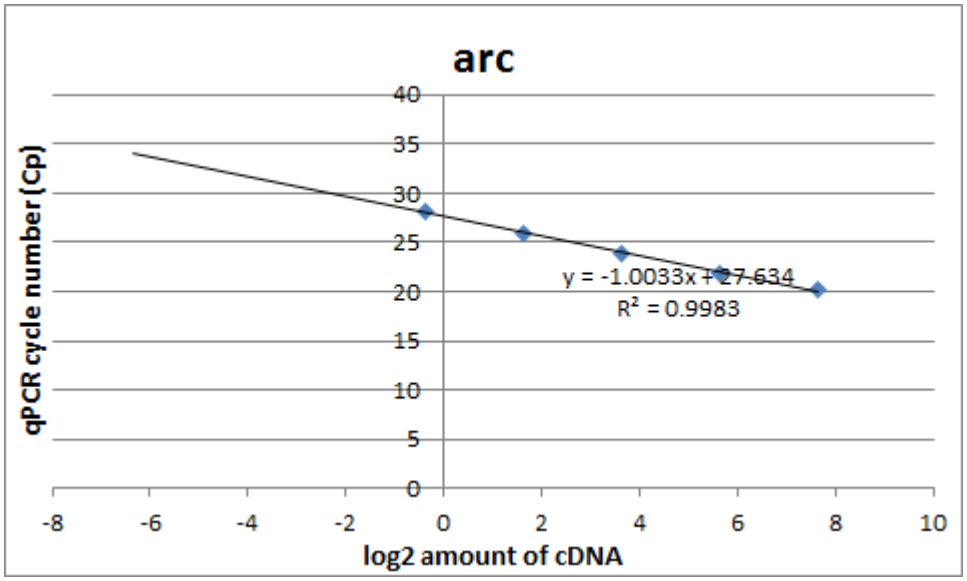
Table A2.1 – Primers used for qPCR

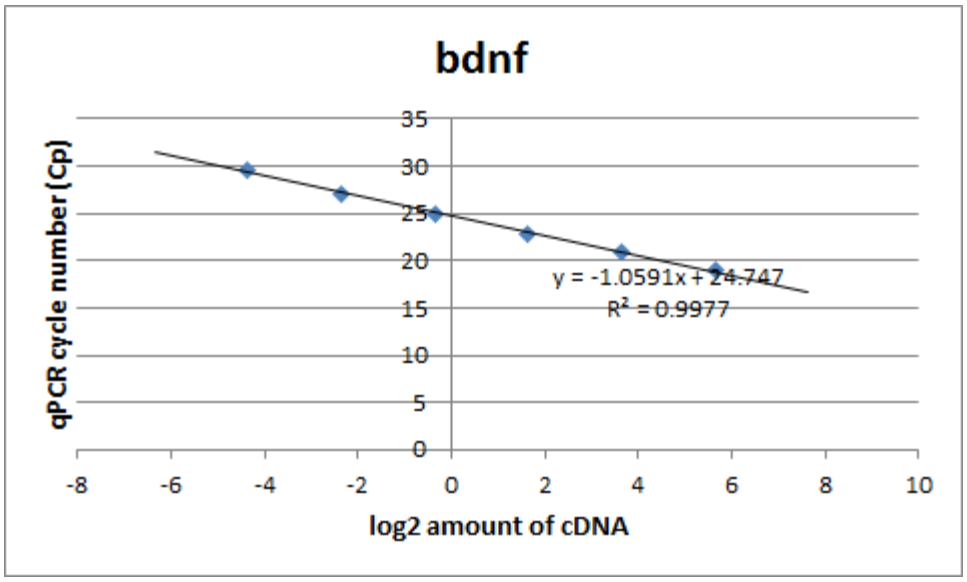
FWD_ACTIN	CTG GCC GGG ACC TGA CAG ACT ACC
RC_ACTIN	TCT TTG ATG TCA CGC ACG ATT TCC CT
FWD_ARC	CGC AGA AGC AGG GTG AAC CAC TCG
RC_ARC	GCA GAA AGC GCT TGA GTT TGG GCT G
FWD_BDNF	AGA AAG TCC CGG TAT CCA AAG GCC
RC_BDNF	ATT GGG TAG TTC GGC ATT GCG AGT
FWD_CREB1	AGT GCT TGA AAA CCA AAA CAA AAC
RC-CREB1	ATC TGA TTT GTG GCA GTA AAG GTC
FWD_ERG1	GGC CAA GGC CGA GAT GCA ATT GAT GT
RC-ERG1	AGC CCC GTT GCT CAG CAG CAT CAT CT
FWD_ERK1(MAPK3)	ACC TAC TGT CAG CGC ACG CTG AGG
RC_ERK1(MAPK3)	ACA TTC TCA TGG CGG AAT CGC AGC
FWD_ERK2(MAPK1)	GTC CAT TGA TAT TTG GTC TGT GGG CT

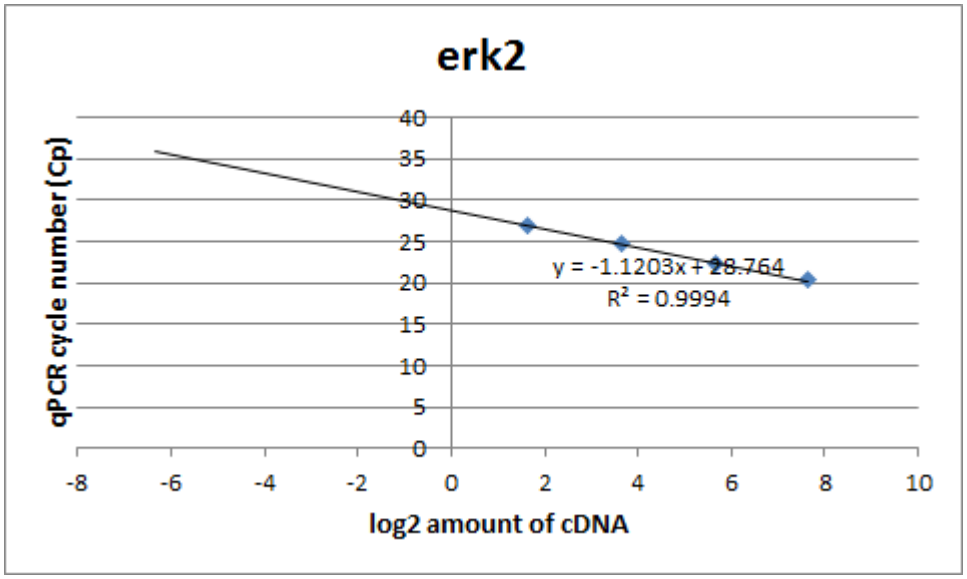
Figure A2.3: Standard curves of primers used in qPCR

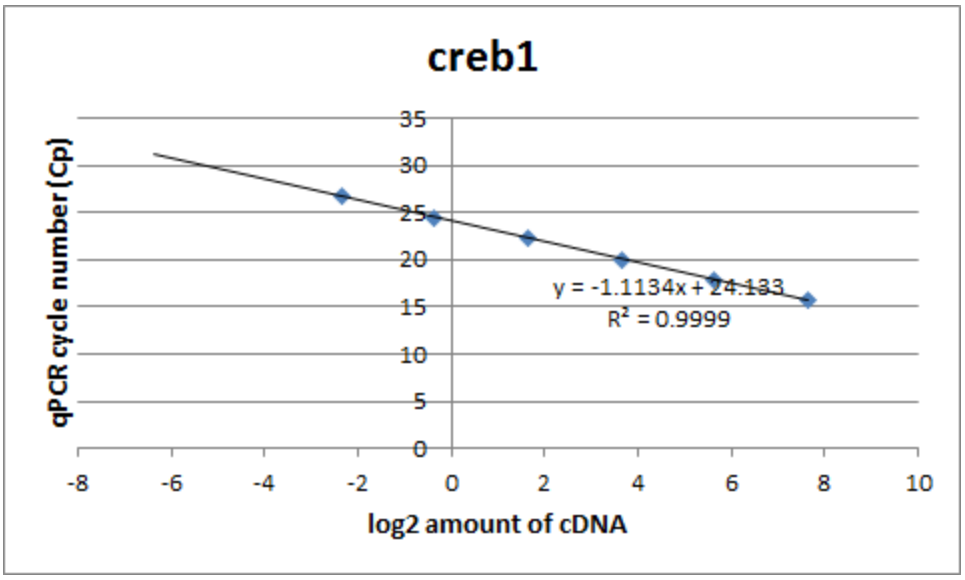
The observed Cp values are plotted against the eight 4-fold dilutions of cortex cDNA for each of the primer pairs used in the study. The cDNA amount in ng is represented on a $\log_2$ scale. Each point on the plot represents a particular dilution The slope of the line is representative of the efficiency of the primers where a slope of -1 indicates 100% efficiency and values more negative than -1 indicate reduced efficiency.

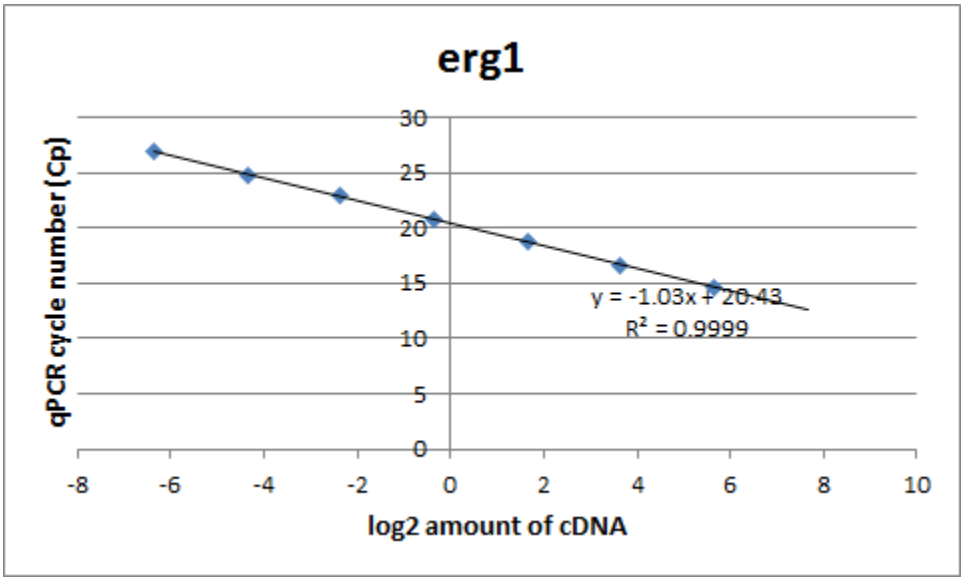

act1
qPCR cycle number (Cp)
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
y = -1.0547x + 21.703
R² = 0.9997
log2 amount of cDNA

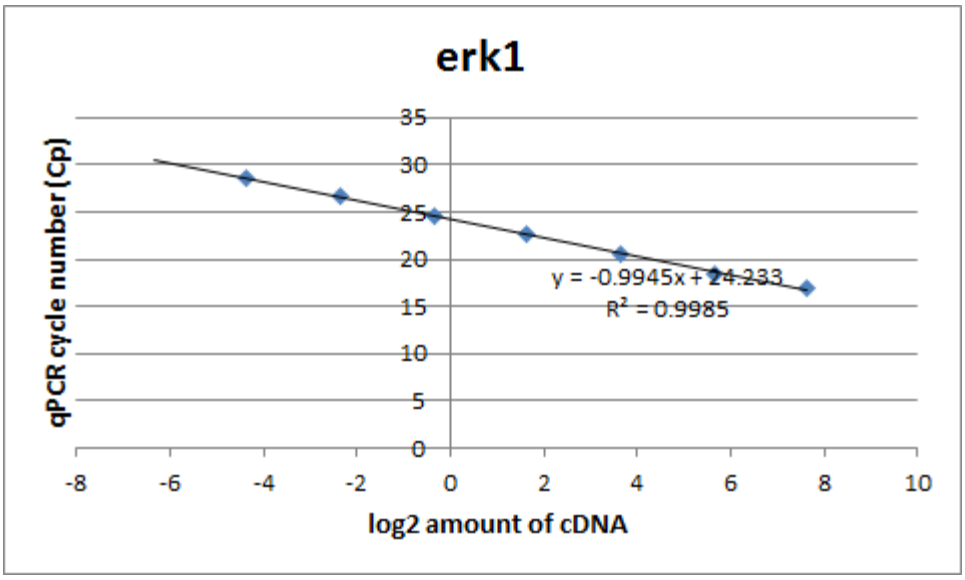

arc
qPCR cycle number (Cp)
40
35
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
y = -1.0033x + 27.634
R² = 0.9983
log2 amount of cDNA

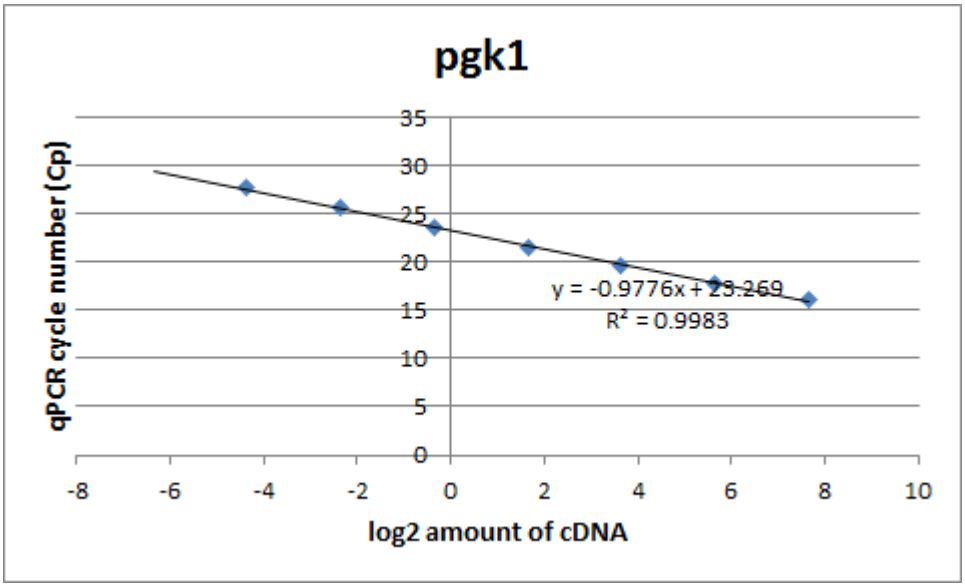

bdnf
qPCR cycle number (Cp)
35
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
log2 amount of cDNA
y = -1.0591x + 24.747
R² = 0.9977

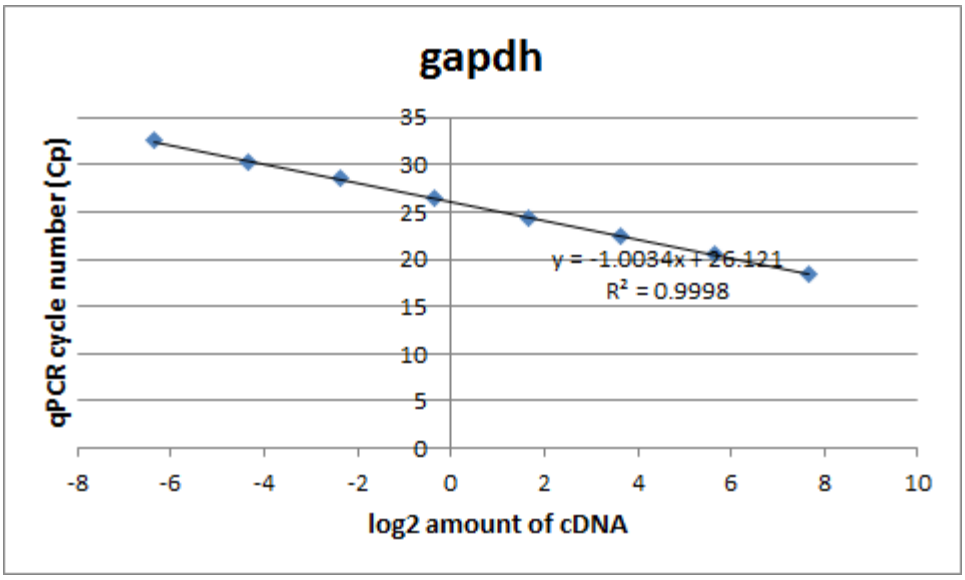

erk2
qPCR cycle number (Cp)
40
35
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
log2 amount of cDNA
y = -1.1203x + 28.764
R² = 0.9994

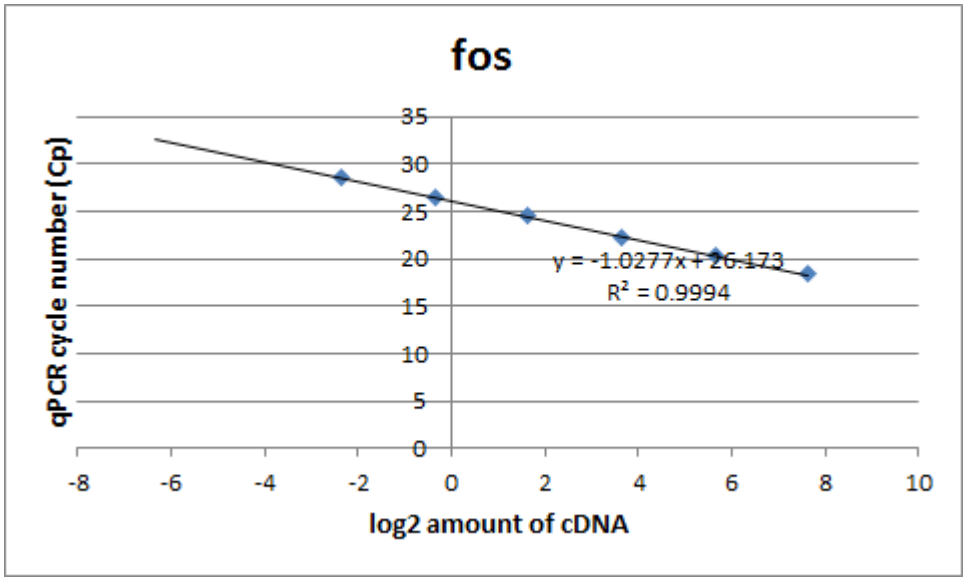

creb1
qPCR cycle number (Cp)
35
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
log2 amount of cDNA
y = -1.1134x + 24.133
R² = 0.9999


erg1
qPCR cycle number (Cp)
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
log2 amount of cDNA
y = -1.03x + 20.43
R² = 0.9999


erk1
qPCR cycle number (Cp)
35
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
log2 amount of cDNA
y = -0.9945x + 24.233
R² = 0.9985


pgk1
qPCR cycle number (Cp)
35
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
log2 amount of cDNA
y = -0.9776x + 23.269
R² = 0.9983


gapdh
qPCR cycle number (Cp)
35
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
log2 amount of cDNA
y = -1.0034x + 26.121
R² = 0.9998


fos
qPCR cycle number (Cp)
35
30
25
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
8
10
log2 amount of cDNA
y = -1.0277x + 26.173
R² = 0.9994

Data analysis

Behavior. In each trial, mice either dug first in the correct (rewarded) odor or failed to do so (i.e., dug first in the nonrewarded odor or, rarely, failed to dig at all). Analyses were performed on the proportion of correct selections across each 20-trial session. A "1" was assigned to each trial in which the mouse made a correct selection, and a "0" otherwise. The scores from blocks of five consecutive trials (*trial blocks*) were averaged together to create a *proportion correct* value for that trial block; accordingly, each 20-trial session comprised four trial blocks.

Because the dependent measures were not continuous, unbounded variables, and thereby violate two assumptions for linear models, we performed a logit transformation prior to statistical analysis. Specifically, we replaced all proportion correct values (X) of 0 with 0.01 and values of 1 with 0.99, and transformed them using the formula ln(X/(1-X)). We then ran a linear mixed effects analysis on the transformed measures with three fixed effects: *Day* (1, 2, 4, or 6), *Trial Block* (TB1-TB4 within each day), and *Odor* (pentanoic acid or butanoic acid). *Mouse* was included as a random effect to control variance arising from individual differences. We used estimated marginal means with Bonferroni correction to perform pairwise comparisons on significant interactions from the full model. All analyses were performed using IBM SPSS 23.0.

mRNA expression levels. We were interested in the changes in mRNA transcript levels for each gene evoked by incremental learning within each brain region within an hour following each day's training session and across the course of six consecutive days of training. We used the amount, normalized to β-Actin levels, as the measure of mRNA transcript levels.

We then used a linear mixed effects model with three fixed effects: *Delay* (five levels: immediately before training, *pre*; immediately after training, *0*; or *15, 30,* or *60* minutes following the end of training), *Region* (five levels: cerebellum, cortex, hippocampus, olfactory bulb, striatum), and *Day* (four levels: day 1, 2, 4, or 6). *Mouse* and *region nested within mouse* were included as random effects. We ran a separate model for each gene of interest. We used estimated marginal means with Bonferroni correction to perform pairwise comparisons on significant interactions from the full model. All analyses were performed using IBM SPSS 23.0.

RESULTS

Progressive learning of the odor-reward association

We first analyzed behavioral performance trajectories across all days of training to assess patterns of learning. We ran a linear mixed model with three fixed effects, *Day* (1, 2, 4, or 6), *Trial Block* (TB1-TB4 within each day), and *Odor* (pentanoic acid or butanoic acid); *Mouse* was included as a random effect. The main effect of Odor was not significant ($F(1, 42.225) = .575$, $p = .452$) indicating that performance was the same for animals regardless of which odor was rewarded. We observed a significant main effect of Day ($F(3, 622.114) = 11.7835$, $p < .001$) as well as a significant main effect of Trial Block ($F(3, 622.426) = 10.435$, $p < .001$), although their interaction was not significant ($F(3, 622.426) = 1.034$, $p = .418$).

Post hoc pairwise comparisons (estimated marginal means, using the Bonferroni adjustment for multiple comparisons) showed that the proportion correct performance scores

at the start of each day's training (TB1) were significantly higher on Day 3 and Day 4 than on Day 1 ($p < 0.001$ for all comparisons), indicating that the mice had consolidated a significant amount of odor learning after two days of training (Figure A2.2). Notably, TB1 scores on Days 5 and 6 were not significantly higher than TB1 on Day 1 ($p > 0.05$ for both comparisons), probably because of power loss owing to the substantial reduction in animal numbers after Day 4. Consistent with this interpretation, TB1 performance scores on Days 5-6 remained at essentially the same levels as on Days 3-4. TB1 performance levels did not change significantly after Day 3 ($p > 0.05$ for all comparisons), indicating that cumulative, consolidated odor learning had reached its asymptote on Day 3.

Over the 20 trials (4 trial blocks) that comprised each day of training, the proportion correct performance scores during TB2, TB3, and TB4 all were significantly higher than during TB1 on the first day of training ($p < 0.05$ for all comparisons). On Day 2 of training, only TB3 and TB4 were significantly higher than TB1 ($p < 0.05$). On subsequent training days, none of the trial blocks were significantly higher than TB1 for that day ($p > 0.05$ for all comparisons). These results corroborate the finding that odor discrimination learning had reached its asymptote on Day 3, and further indicate that additional training on or after Day 3 did not significantly improve performance over that enabled by previously consolidated learning (Figure A2.2).

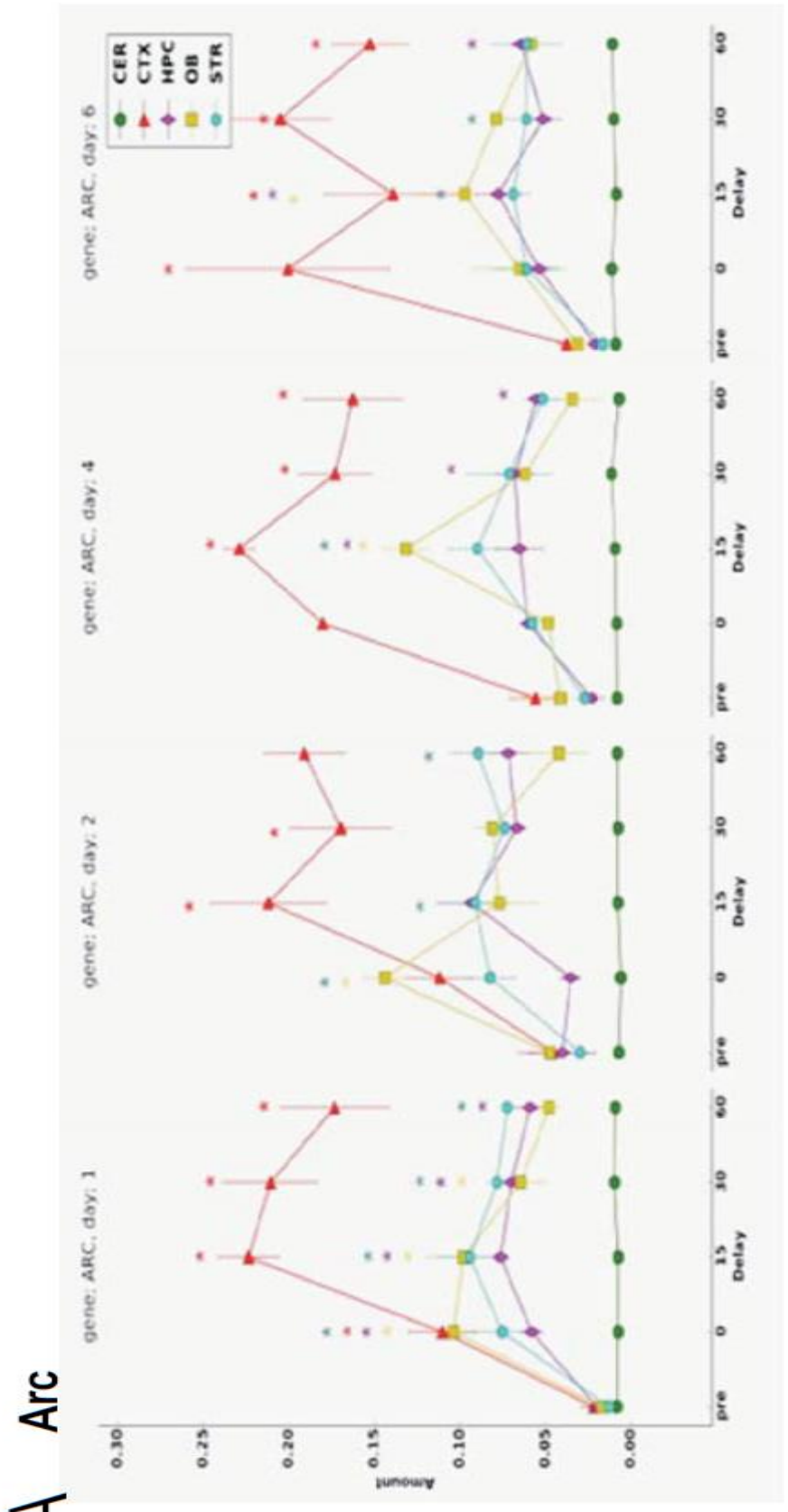
Learning-induced transcription of immediate-early genes

We measured the expression of the immediate-early genes (IEGs) Arc, Egr1 (aka Zif268, NGFI-A, Krox24), and Fos at each of the 20 time points at which brain tissue was sampled (Figure A2.4), and assessed the significance of changes in mRNA expression levels across

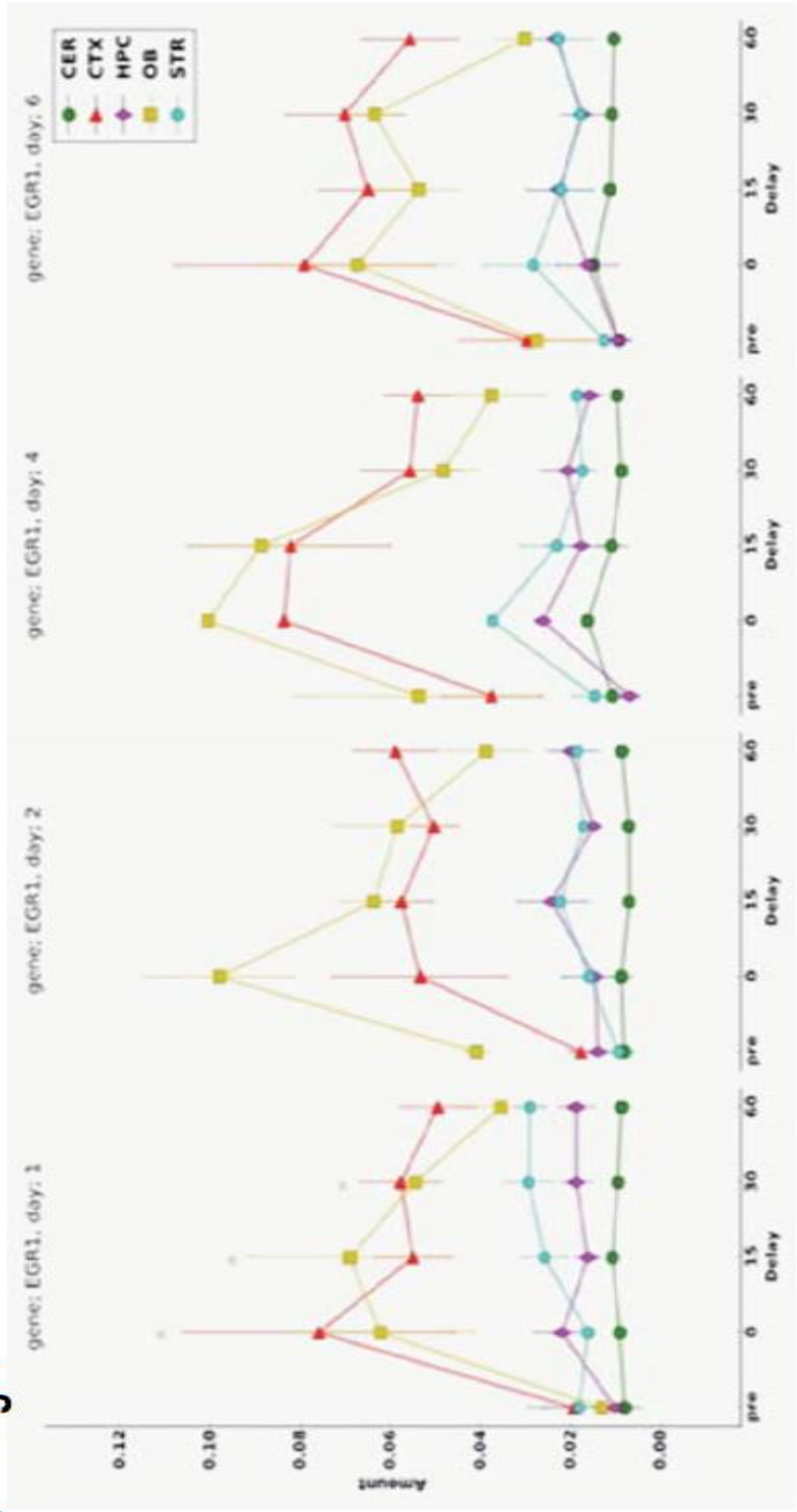
the period from the start of learning until 60 minutes after the end of learning on a given day (*Delay*), across a selected four of the six successive days of learning (*Day*), and across the five regions of the brain sampled (*Region*).

For Arc, our linear mixed effects model indicated significant main effects of Delay ($F(4, 35.734) = 40.356$, $p < 0.001$) and Region ($F(4, 138.180) = 363.846$, $p < 0.001$), but not Day ($p > 0.05$). The main effect of Region indicates that average Arc mRNA levels differed between the five brain regions sampled (Figure A2.4A), whereas the main effect of Delay indicates that mRNA levels changed over the course of learning on a given day. There also was a significant two-way interaction of Region and Delay ($F(16, 138.479) = 5.595$, $p <.001$), indicating that the timecourse of Arc expression levels over the course of learning

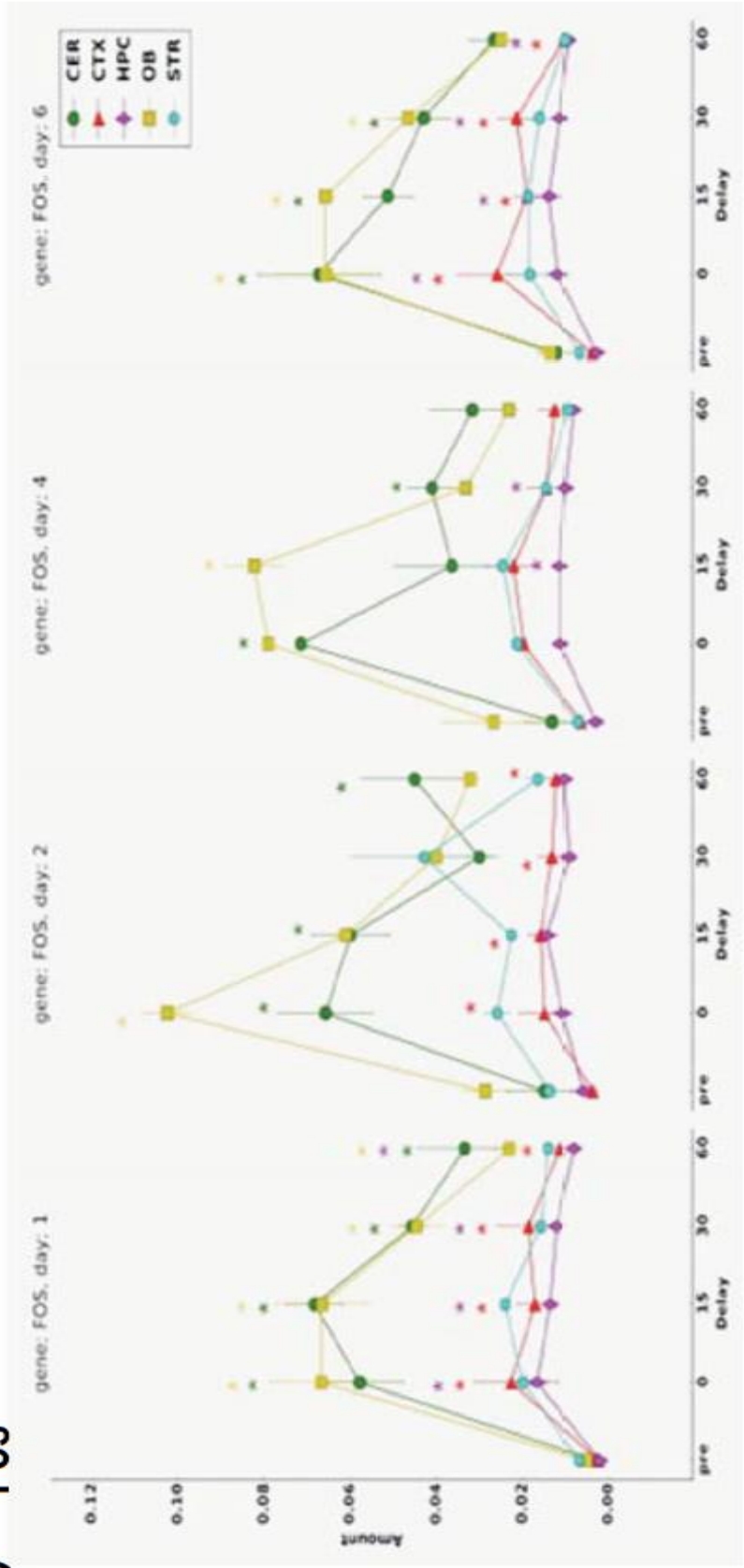
Figure A2.4: (A-C) - The normalized amount for each transcript is plotted for five time blocks for each day across five brain regions. Each point is derived from upto three biological replicates with three technical replicates for each. (D) - Comparison of Day 1 data across the three genes arc, egr1 and fos.


A
Arc
gene: ARC, day: 1
gene: ARC, day: 2
gene: ARC, day: 4
gene: ARC, day: 6
CER
CTX
HPC
OB
STR
Amount
0.30
0.25
0.20
0.15
0.10
0.05
0.00
pre
0
15
30
60
Delay

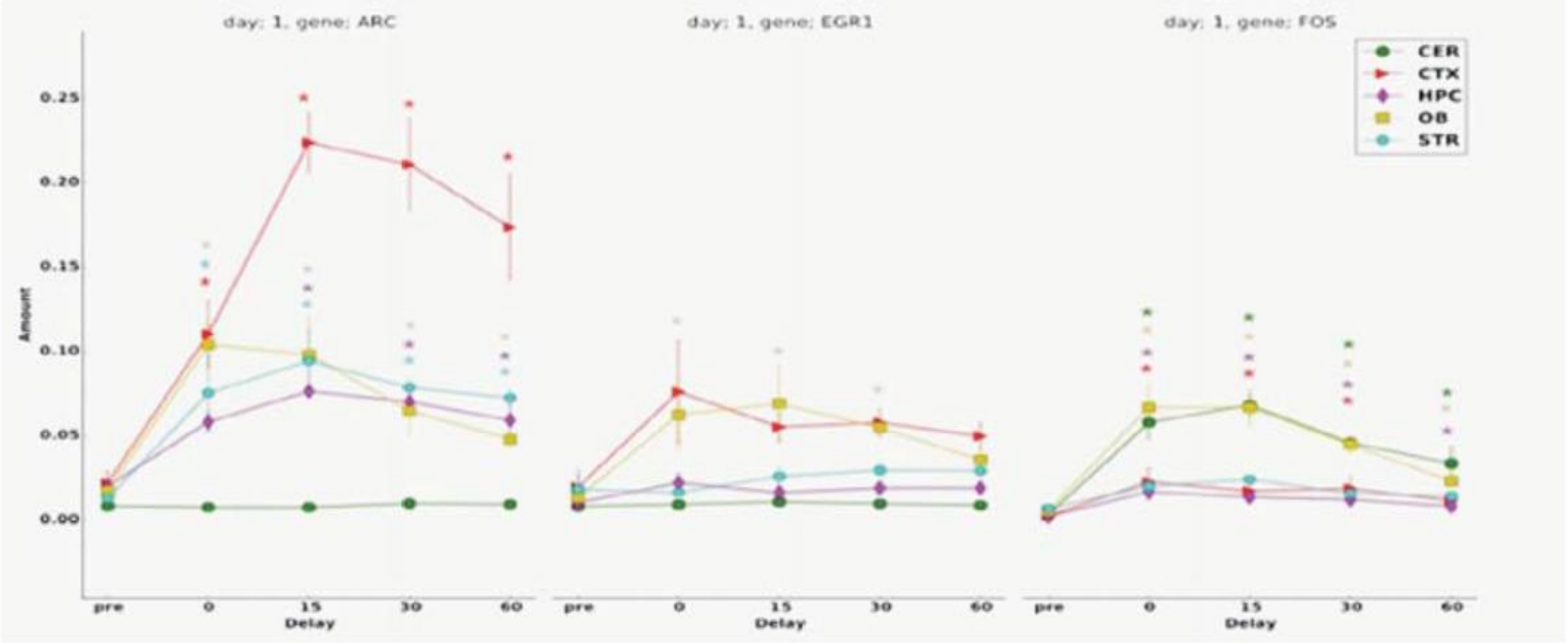
B Egr1


gene: EGR1, day: 1
gene: EGR1, day: 2
gene: EGR1, day: 4
gene: EGR1, day: 6
CER
CTX
HPC
OB
STR
Amount
0.12
0.10
0.08
0.06
0.04
0.02
0.00
pre
0
15
30
60
Delay

C Fos


gene: FOS, day: 1
gene: FOS, day: 2
gene: FOS, day: 4
gene: FOS, day: 6
CER
CTX
HPC
OB
STR
Amount
0.12
0.10
0.08
0.06
0.04
0.02
0.00
pre
0
15
30
60
Delay

D Quantitative comparison of IEGs



differed among the five regions sampled. There were no other significant main effects or interactions. Notably, the absence of a significant effect of Day suggested that the expression of Arc mRNA induced by learning did not change as a result of the consolidated learning accumulated over multiple days of training (Figure A2.4A).

Post hoc pairwise comparisons revealed that Arc mRNA levels increased significantly above baseline (*pre*) in most brain regions and at most delays after odor discrimination learning on each day of training (Figure A2.4A; *asterisks* indicate $p < 0.05$ compared with timepoint *pre*). Only in the cerebellum was there no Arc response to odor learning. Together, these results show that Arc mRNA levels increased robustly after incremental, appetitive odor learning on each day in isocortex, hippocampus, olfactory bulb, and striatum, though not to the same extent across these regions, and that the timecourse of expression did not differ across subsequent days of training.

We performed the same analyses on Egr1 mRNA expression data. For Egr1, our analysis again yielded significant main effects of Delay ($F(4, 36.859) = 7.216$, $p < 0.001$) and Region ($F(4, 138.040) = 152.939$, $p < 0.001$), but not Day ($p > 0.05$), as well as a significant two-way interaction of Region and Delay ($F(16, 138.106) = 2.052$, $p = 0.014$). There were no other significant main effects or interactions. The profile of these results was qualitatively identical to the response of Arc, inviting a similar interpretation; however, the pattern of activation responses across brain regions differed. In particular, the transcriptional activation of Egr1 was as strong in the olfactory bulb as in isocortex, whereas this was not the case for Arc (Figure A2.4B).

Finally, we performed the same analyses on Fos data. Analysis of Fos mRNA expression levels again showed main effects of Delay ($F(4, 32.133) = 38.221$, $p < 0.001$) and Region

($F(4, 131.066) = 237.798$, $p < 0.001$), but not Day ($p > 0.05$), as well as a significant two-way interaction of Region and Delay ($F(16, 130.931) = 3.337$, $p < 0.001$). There were no other significant main effects or interactions. The profile of these results was qualitatively identical to the responses of Arc and Egr1, again inviting a similar interpretation. However, the pattern of activation responses across brain regions again differed from those of the other two IEGs; notably, there was a strong response from cerebellum, as well as olfactory bulb, and the isocortical response was weaker than with the other IEGs tested (Figure A2.4C; significant increases in expression compared to time point *pre* are indicated with *asterisks*).

For comparison, Day 1 mRNA transcriptional activation profiles for the three IEGs tested are depicted on the same absolute scale in Figure A2.4D.

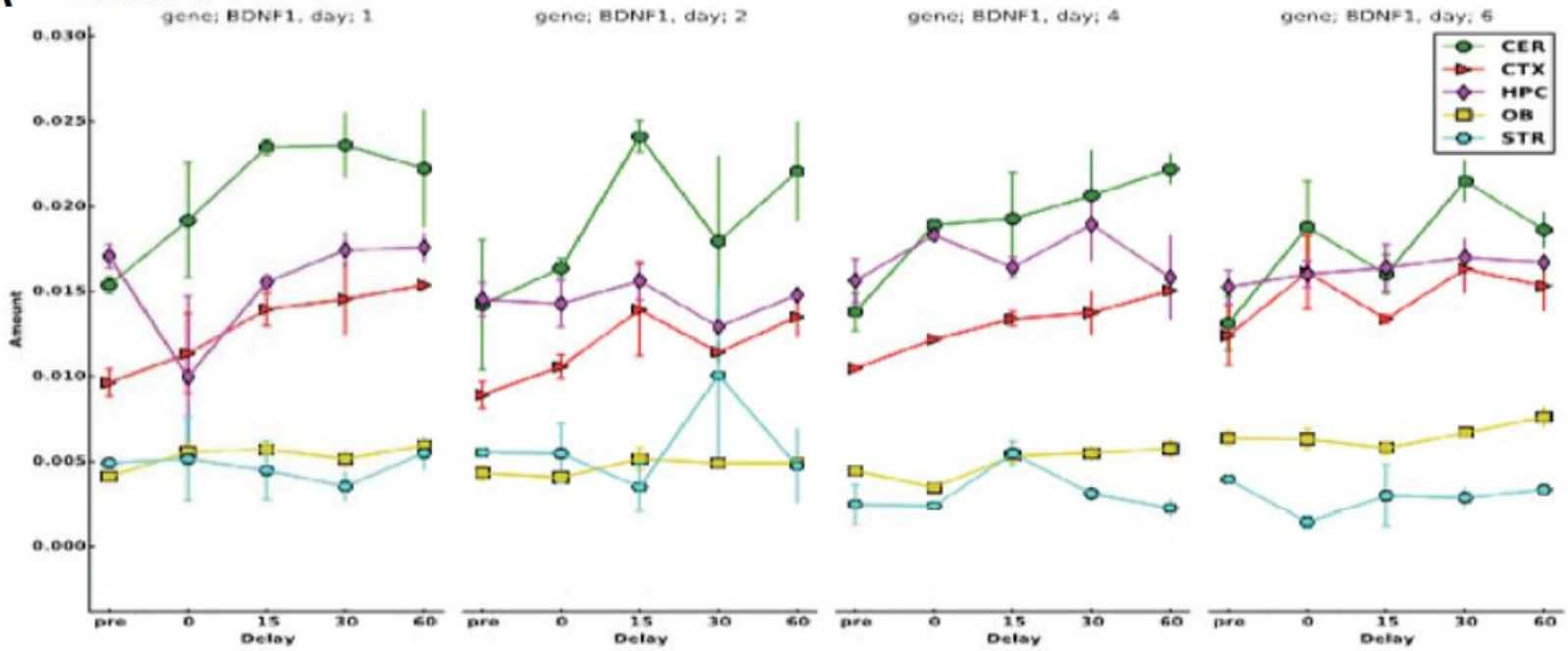
Brain-derived neurotrophic factor (BDNF) mRNA levels differ across brain regions

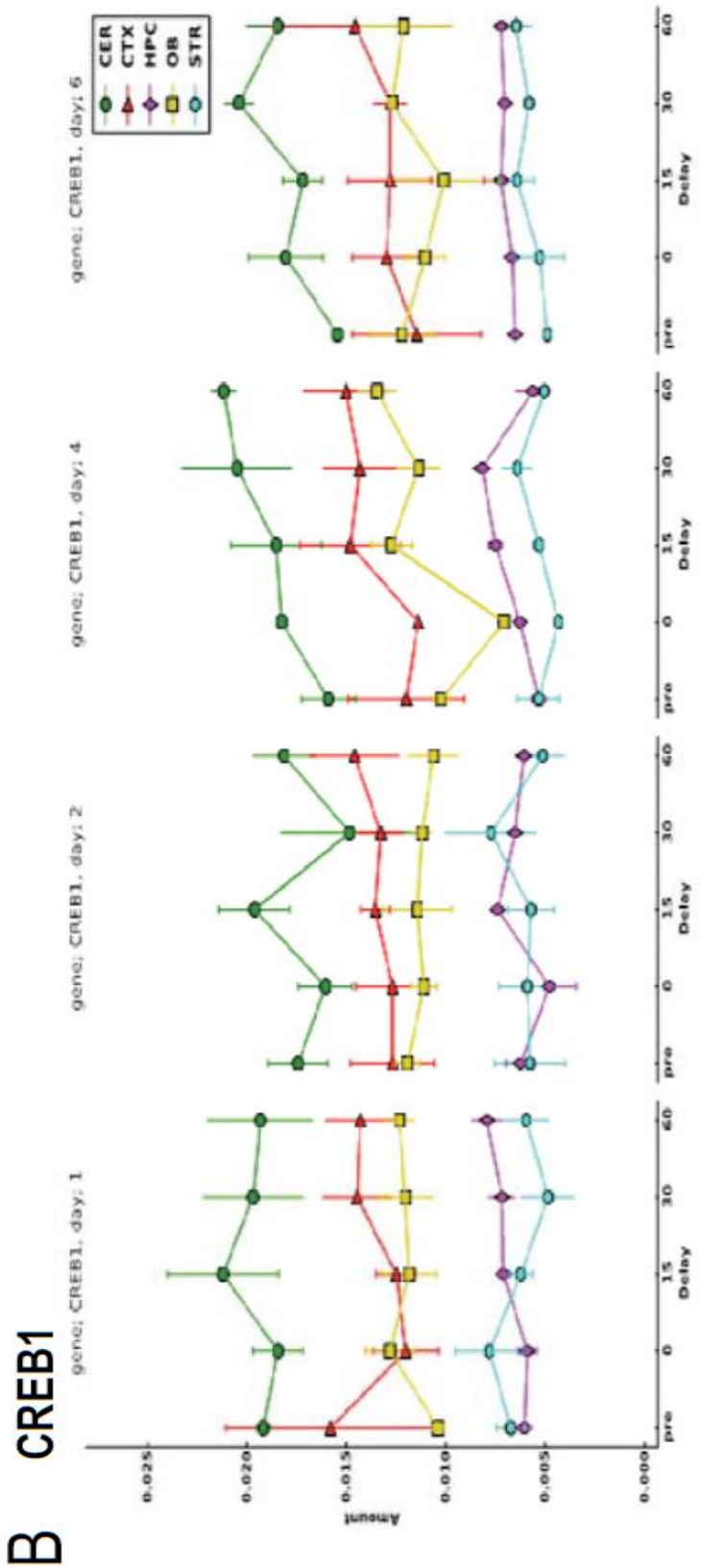
We also compared mRNA transcriptional activation for four additional plasticity-related proteins induced by odor discrimination learning: brain-derived neurotrophic factor (BDNF), the transcription factor cyclic AMP response element binding protein (CREB), and two extracellular signal-related kinases (ERK1, ERK2).

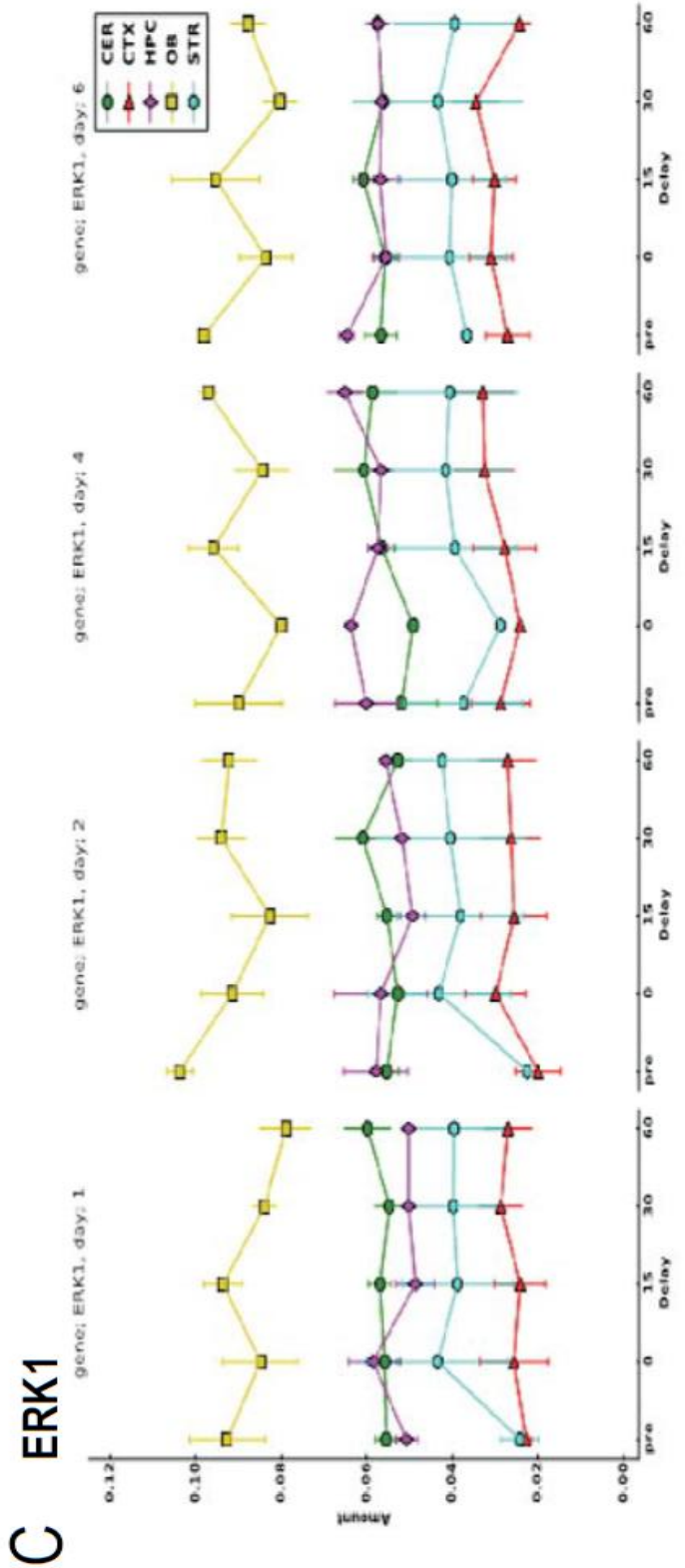
For BDNF, our analyses suggested that the mRNA transcript levels did not increase in response to training on any of the days (Figure A2.5 A). There were no significant effects of Delay or Day. However, the levels showed a significant effect for Region.

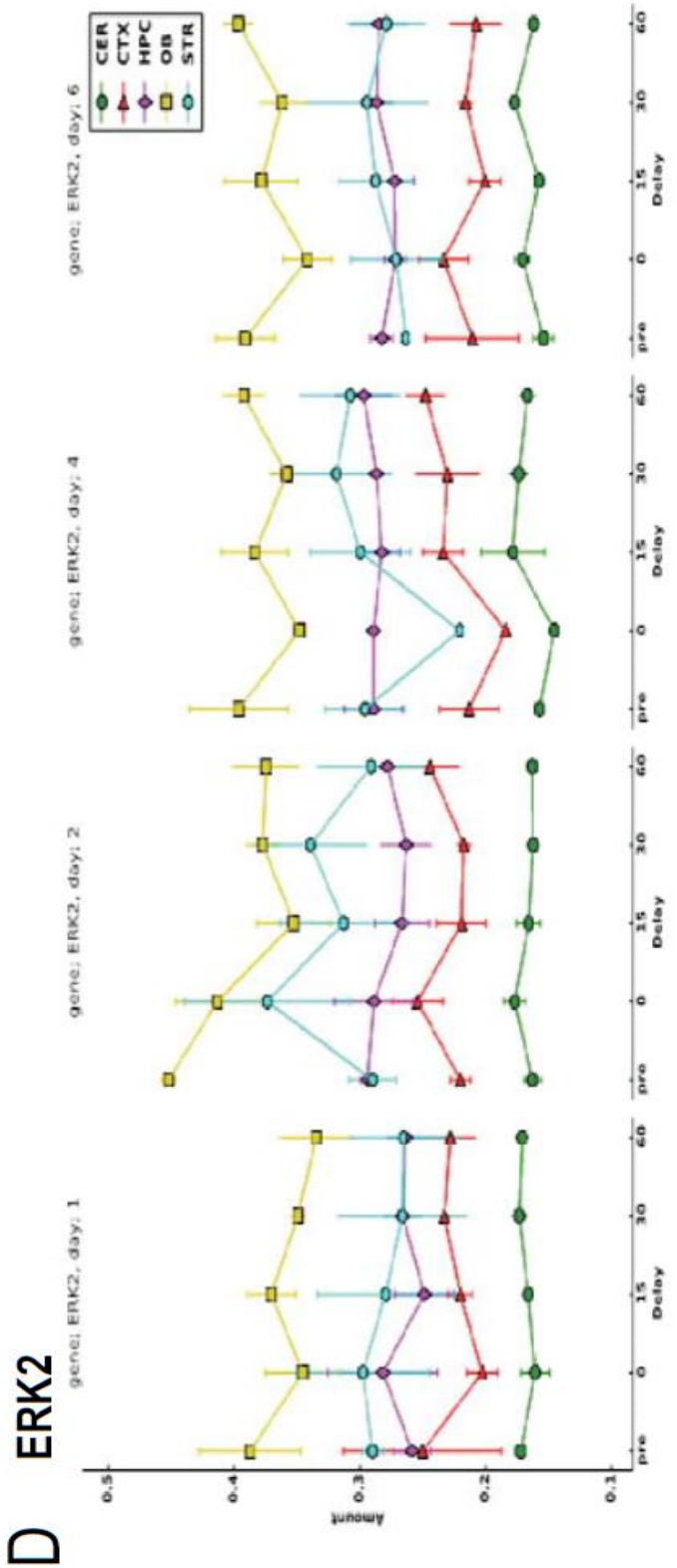
Figure A2.5: (A-D) - The normalized amounts for each transcript are plotted for five time blocks for each day across five brain regions. Each point is derived from upto three biological replicates with three technical replicates for each.

A BDNF1


gene; BDNF1, day; 1
gene; BDNF1, day; 2
gene; BDNF1, day; 4
gene; BDNF1, day; 6
CER
CTX
HPC
OB
STR
Amount
0.030
0.025
0.020
0.015
0.010
0.005
0.000
pre
0
15
30
60
Delay


B
CREB1
gene; CREB1, day; 1
gene; CREB1, day; 2
gene; CREB1, day; 4
gene; CREB1, day; 6
CER
CTX
HPC
OB
STR
Amount
0.025
0.020
0.015
0.010
0.005
0.000
pre
0
15
30
60
Delay


C
ERK1
gene; ERK1, day; 1
gene; ERK1, day; 2
gene; ERK1, day; 4
gene; ERK1, day; 6
CER
CTX
HPC
OB
STR
Amount
0.12
0.10
0.08
0.06
0.04
0.02
0.00
pre
0
15
30
60
Delay


D
ERK2
gene; ERK2, day; 1
gene; ERK2, day; 2
gene; ERK2, day; 4
gene; ERK2, day; 6
CER
CTX
HPC
OB
STR
Amount
0.5
0.4
0.3
0.2
0.1
pre
0
15
30
60
Delay

Odor discrimination training does not significantly activate transcription of CREB1, Erk1, or Erk2

For CREB1, analysis with our linear mixed effects model revealed only a main effect of Region ($F(4, 138.642) = 160.301$, $p < 0.001$), indicating that the levels of CREB mRNA expression differed among the brain regions studied. There were no other significant main effects or interactions; in particular, the absence of a significant effect of Delay or Day indicated that CREB1 mRNA levels did not change significantly in response to odor discrimination learning, either on a given day or across the 6 days of training (Figure A2.5B).

For ERK1, statistical analysis revealed only a main effect of Region ($F(4, 139.228) = 115.764$, $p < 0.001$), indicating that the levels of ERK1 mRNA expression differed among the brain regions studied. There were no other significant main effects or interactions; in particular, the absence of a significant effect of Delay or Day indicated that ERK1 mRNA levels did not change significantly in response to odor discrimination learning, either on a given day or across the 6 days of training (Figure A2.5C).

For ERK2, statistical analysis revealed only a main effect of Region ($F(4, 173.000) = 191.291$, $p < 0.001$), indicating that the levels of ERK2 mRNA expression differed among the brain regions studied. There were no other significant main effects or interactions; in particular, the absence of a significant effect of Delay or Day indicated that ERK2 mRNA levels did not change significantly in response to odor discrimination learning, either on a given day or across the 6 days of training (Figure A2.5D).

DISCUSSION

In the current study, we trained animals on an associative, appetitive odor discrimination task over the course of 6 days. We found that animals were able to acquire this association and reach asymptotic performance by Day 3 of training when presented with new odor cues. Using high-throughput RT-PCR on samples from five separate brain regions, we measured mRNA levels before training and at four time points following training on four of the six days of training. We found that all eight of the genes tested were transcribed at significantly different levels across the five brain regions sampled, and that their relative expression levels, with respect to one another, also differed substantially among these regions. The three IEGs tested were strongly transcriptionally activated by training, with mRNA levels rising sharply after training (in a gene-specific selection of brain areas) and falling back to baseline on an hour(s) timescale. However, there was no change in the IEG response profiles over the six successive days of conditioning, despite the clear behavioral evidence of cumulative learning and memory consolidation over the first three days. The plasticity-related protein transcripts, CREB1, Erk1, BDNF and Erk2, were not responsive to odor discrimination learning on either the minutes or days timescale.

Region-specificity

We know from many studies that sub-region/cell-type specific changes can differ greatly in level, but also in time course. For example, in the nucleus accumbens, following a single cocaine injection, BDNF mRNA levels peak 1 hour after injection, but in the shell only, and not the core (Graham et al., 2007). Following a fear-potentiated startle paradigm, where the

strength of a startle response to an acoustic stimulus is elevated in the presence of light that has been previously paired with a foot shock, BDNF mRNA levels were high in the basal lateral amygdala two hours following fear conditioning and returned to baseline levels by four hours after training. No changes in BDNF mRNA levels were observed in the HPC, medial nucleus of the amygdala or ventral posteromedial nucleus of the thalamus (Rattiner et al., 2005) suggesting that endogenous BDNF activity in response to learning events is region-specific.

Immediate-early genes

As expected, we found that the immediate-early genes Arc, Egr1, and Fos, each of which has been used extensively as a marker of learning-responsive cellular activation, rapidly increased their mRNA expression levels immediately following associative odor learning. We also found differences both in the absolute levels of expression of the three genes across the five brain regions studied (Figure A2.4D) and in the relative levels of expression measured in these different regions. For example, Arc was most strongly activated in isocortex, only modestly activated in the hippocampus, olfactory bulb, and striatum, and not at all in cerebellum. Egr1 was activated comparably strongly in isocortex and olfactory bulb, and only marginally or not at all in the other three regions. Fos was activated most strongly in the olfactory bulb and cerebellum, and marginally or not at all in isocortex, hippocampus, and striatum.

BDNF

Blockade of BDNF receptor (and other kinase) activity in the OB by infusion of the TrkB

inhibitor K252a during learning significantly impairs long-term (48 hr) memory for the associative odor discrimination task used in this study. This finding is consistent with the well-established effects of BDNF signaling blockade on multiple forms of memory, but additionally demonstrates the essential role of olfactory bulb circuitry in the consolidation of associative odor discrimination memory in the context of this task. Accordingly, we anticipated that BDNF expression levels in olfactory bulb would be likely to increase following odor discrimination training.

In fact, BDNF transcript levels in olfactory bulb were unchanged following training either on the minutes or days timescale. Either of two factors may resolve this apparent inconsistency. First, it is likely that BDNF secretion in the olfactory bulb is mediated primarily by extrabulbar neurons and transported along their axons for release into the bulb. Under these circumstances, BDNF mRNA would not be strongly expressed within the olfactory bulb at any point, despite the substantial release of BDNF peptide therein. Indeed, total BDNF mRNA levels in olfactory bulb have been previously shown to be substantially lower than those in the hippocampus (Malkovska et al., 2006). Second, the BDNF peptide is translated from as many as 24 different mRNA transcripts based on different combinations of nine promoters and alternative splicing of multiple exons, evincing a rich and intricate set of transcriptional control mechanisms regulating BDNF synthesis. The primers used here target exon 9 of BDNF, a constitutive exon, and as such they cannot distinguish between alternative isoforms of BDNF. An alternative approach using primers targeting all the exon junctions in the 24 individual isoforms and coupled to deep sequencing (Larson et al., 2016) could be used to quantify the changes in abundance of the alternative exon junctions and determine if they change with learning.

Other plasticity-related proteins

We found that transcript levels for CREB1 (coding for the transcription factor CREB) and ERK1 and ERK2 (coding for the intracellular signaling genes Erk1 and Erk2) are expressed at different levels in different regions of the brain, but do not significantly change in response to learning. Multiple studies of the regulatory mechanisms associated with these molecules have established that they are often important for establishment of long-term memory (Shaywitz & Greenberg, 1999; reviewed by Tong et al., 2014). While the work presented here suggests that overall levels of these transcripts are unchanged during learning, post-transcriptional regulatory processes like splicing or translational regulation of these genes could play a role in consolidation of long-term memory.

Summary

In summary, the strength of the current study is that it generated a "spatiotemporal map" of plasticity-related mechanisms following multi-trial learning. Future studies can use this "map" to more specifically probe the role of these mechanisms, as well as the importance of their timing. For example, we can ask questions about the functional role of Fos activity in the OB at a given time after learning, including behavioral and structural effects. In the present study, we observed increases in Fos mRNA in the OB immediately and 15 minutes after multi-trial training on all the days. Is this brief time frame crucial for LTM consolidation? We can test this by artificially elevating Fos levels in the OB following learning beyond 15-minutes and then using behavioral testing to assess the effect of this manipulation on LTM. More impactful, however, is the fact the map captures the time courses of multiple brain regions for the same learning event. This allows us to test

hypotheses about how brain regions may coordinate with each other for higher-order computations.

Apart from the transcripts targeted for investigation here, a genome-wide investigation of changes in expression would provide powerful insights into the molecular mechanisms of learning. While such approaches were cost-prohibitive when this study began, recent technological improvements and innovations have made such an approach plausible. While a traditional RNA-seq experiment would be very expensive on a set of 300 samples, newer approaches like Quant-seq (Moll et. al, 2014) can be used to sequence the 3' end of transcripts and quantify their abundances. Though they would not be able to discriminate between the different splicing isoforms, they would still provide insightful information on overall transcript levels. This information can then be used to identify targets which respond to multi-trial learning for more detailed investigation.

REFERENCES

Tamara Aid, Anna Kazantseva, Marko Piirsoo, Kaia Palm, and Tonis Timmusk. Mouse and rat *bdnf* gene structure and expression revisited. *Journal of Neuroscience Research*, 85(3):525–535, 2007.

Mariana Alonso, Pedro Bekinschtein, Martin Cammarota, Monica RM Vianna, Ivan Izquierdo, and Jorge H Medina. Endogenous BDNF is required for long-term memory formation in the rat parietal cortex. *Learning and Memory*, 12(5):504–510, 2005.

Pedro Bekinschtein, Cammarota Martin, LM Igaz, LRM Bevilaqua, Izquierdo Ivan, and Medina Jorge. Persistence of long-term memory storage requires a late protein synthesis and BDNF-dependent phase in the hippocampus. *Neuron*, 53(2):261–277, 2007.

Ramon Bernabeu, Lia Bevilaqua, Patricia Ardenghi, Elke Bromberg, Paulo Schmitz, Marino Bianchin, Ivan Izquierdo, and Jorge H Medina. Involvement of hippocampal D1/D5 receptor-cAMP signaling pathways in a late memory consolidation phase of an aversively-motivated task in rats. *Proceedings of the National Academy of Sciences of the United States of America*, 94(13):7041–7046, 1997a.

Ramon Bernabeu, Martin Cammarota, Ivan Izquierdo, and Jorge H Medina. Involvement of glutamate AMPA receptors and a cAMP/protein kinase A/CREB-P pathway in memory consolidation of an aversive learning task in rats. *Brazilian Journal of Medical and Biological Research*, 30(8):961–965, 1997b.

Ramon Bernabeu, P Schmitz, M P Faillace, Ivan Izquierdo, and Jorge H Medina. Hippocampal cGMP and cAMP are differentially involved in memory processing of an inhibitory avoidance learning. *NeuroReport*, 7:585–588, 1996.

L Bevilaqua, Patricia Ardenghi, N Schroder, E Bromberg, PK Schmitz, E Schaeffer, J Quevedo, M Bianchin, R Walz, Jorge H Medina, and Ivan Izquierdo. Drugs acting upon the cyclic adenosine monophosphate/protein kinase A signaling pathway modulate memory consolidation when given late after training into rat hippocampus but not amygdala. *Behavioural Pharmacology*, 8(4):331–338, 1997.

John F Bishop, Gregory P Mueller, and M Maral Mouradian. Alternate 5 exons in the rat brain-derived neurotrophic factor gene: differential patterns of expression across brain regions. *Molecular Brain Research*, 26(1):225–232, 1994.

Martin Cammarota, G Paratcha, M Levi de Stein, Ramon Bernabeu, Ivan Izquierdo, and Jorge H Medina. B50/GAP43 phosphorylation and PKC activity are increased in rat hippocampal synaptosomal membranes after an inhibitory avoidance learning. *Neurochemical Research*, 22:499–505, 1997.

Dipesh Chaudhury, Laura Manella, Adolfo Arellanos, Olga Escanilla, Thomas A Cleland, and Christiane Linster. Olfactory bulb habituation to odor stimuli. *Behavioral Neuroscience*, 124(4):490, 2010.

Thomas A Cleland, Alix Morse, Esther L Yue, and Christiane Linster. Behavioral models of odor similarity. *Behavioral Neuroscience*, 116(2):222–231, 2002.

Carla Cunha, Riccardo Brambilla, and Kerrie L Thomas. A simple role for BDNF in learning and memory? *Frontiers in Molecular Neuroscience*, 3, 2010.

W Danysz, W Zajaczkowski, and Chris G Parsons. Modulation of learning processes by ionotropic glutamate receptor ligands. *Behavioural Pharmacology*, 6(5):455–474, 1995.

Patrik Ernfors, Cynthia Wetmore, Lars Olson, and Hkan Persson. Identification of cells in rat brain and peripheral tissues expressing mRNA for members of the nerve growth factor family. *Neuron*, 5(4):511–526, 1990

I Ferrer, R Blanco, R River, M Carmona, J Ballabriga, M Olive, and A M Planas. CREB-1 and CREB-2 immunoreactivity in the rat brain. *Brain Research*, 712(1):159–164, 1996.

Danielle L Graham, Scott Edwards, Ryan K Bachtell, Ralph J DiLeone, Maribel Rios, and David W Self. Dynamic BDNF activity in nucleus accumbens with cocaine use increases self-administration and relapse. *Nature Neuroscience*, 10(8):1029–1037, 2007

Ivan Izquierdo, C Da Cunha, R Rosat, D Jerusalinsky, MBC Ferreira, and Jorge H Medina. Neurotransmitter receptors involved in memory processing by the amygdala, medial septum and hippocampus of rats. *Behavioral and Neural Biology*, 58:16–25, 1992.

Luciana A Izquierdo, N Schroder, Patricia Ardenghi, J Quevedo, L Bevilaqua, CA Netto, Ivan Izquierdo, and Jorge H Medina. Systemic administration of ACTH or vasopressin in rats reverses the amnestic effect of post-training b-endorphin or electroconvulsive shock but not that of intra-hippocampal infusion of protein kinase inhibitors. *Learning and Memory*, 68(2):197–202, 1997.

Diane Jerusalinsky, J H Quillfeldt, R Walz, RC Da Silva, Jorge H Medina, and Ivan Izquierdo. Post-training intrahippocampal infusion of protein kinase C inhibitors causes retrograde amnesia in rats. *Behavioral and Neural Biology*, 61:107–109, 1994.

Diane Jerusalinsky, MBC Ferreira, RC Da Silva, M Bianchin, A Ruschel, Jorge H Medina, and Ivan Izquierdo. Amnesia by infusion of glutamate receptor blockers into the amygdala, hippocampus and entorhinal cortex. *Behavioral and Neural Biology*, 58:76–80, 1992.

Bonnie E Lonze and David D Ginty. Function and regulation of *creb* family transcription factors in the nervous system. *Neuron*, 35(4):605–623, 2002.

Dharshan Kumaran and Eleanor A Maguire. Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus*, 17(9):735–748, 2007.

Irena Malkovska, Steven G Kernie, and Luis F Parada. Differential expression of the four untranslated *bdnf* exons in the adult mouse brain. *Journal of Neuroscience Research*, 83(2):211–221, 2006.

Serge Marty, Benedikt Berninger, Patrick Carroll, and Hans Thoenen. GABAergic stimulation regulates the phenotype of hippocampal interneurons through the regulation of brain-derived neurotrophic factor. *Neuron*, 16(3):565–570, 1996.

Makoto Mizuno, Kiyofumi Yamada, N Maekawa, K Saito, M Seishima, and T Nabeshima. CREB phosphorylation as a molecular marker of memory processing in the hippocampus for spatial learning. *Behavioral Brain Research*, 133(2):135–141, 2002.

Makoto Mizuno, Kiyofumi Yamada, A Olariu, H Nawa, and T Nabeshima. Involvement of brain-derived neurotrophic factor in spatial memory formation and maintenance in a radial arm maze test in rats. *Journal of Neuroscience*, 20(18):7116–7121, 2000.

Lisa M Rattiner, Michael Davis, and Kerry J Ressler. Brain-derived neurotrophic factor in amygdala-dependent learning. *The Neuroscientist*, 11(4):323–333, 2005.

Gernot Riedel, Giacomo Casabona, and Klaus G Reymann. Inhibition of long-term potentiation in the dentate gyrus of freely moving rats by the metabotropic glutamate receptor antagonist MCPG. *The Journal of Neuroscience*, 15(1):87–98, 1995.

Tonis Timmusk, Kaia Palm, Madis Metsis, Tonu Reintam, Viiu Paalme, Mart Saarma, and Hkan Persson. Multiple promoters direct tissue-specific expression of the rat *bdnf* gene. *Neuron*, 10(3):475–489, 1993.

Adam J Shaywitz and Michael E Greenberg. CREB: a stimulus-induced transcription factor activated by a diverse array of extracellular signals. *Annual review of biochemistry*, 68(1):821–861, 1999.

Claudia Wolfman, Cyntia Fin, Marcelo Dias, Marino Bianchin, Ricardo C Da Silva, Paulo K Schmitz, Jorge H Medina, and Ivan Izquierdo. Intra-hippocampal or intra-amygdala infusion of KN62, a specific inhibitor of Calcium/Calmodulin-dependent protein kinase II, causes retrograde amnesia in the rat. *Behavioral and Neural Biology*, 61(3):203–205, 1994.