# CHARACTERIZATION OF PRE-MRNA DYNAMICS AND STRUCTURE THROUGHOUT SPLICEOSOME ASSEMBLY AND CATALYSIS

by

Matthew Kahlscheuer

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in the University of Michigan
2015

Doctoral Committee:

Professor Nils G. Walter, Chair
Professor Julie Suzanne Biteen
Professor Anna K. Mapp
Professor Patrick O'Brien

**Acknowledgements**

I would like to thank my family, friends, thesis mentor and Walter lab members for their continued support and encouragement throughout my journey towards achieving a PhD. I would first like to thank my graduate advisor, Dr. Nils Walter. Throughout my six years in graduate school, he has always encouraged and challenged me to become a better scientist and thinker. I would not be the scientist I am today without his continued guidance and mentorship for which I will be forever grateful. I would like to thank my committee members: Dr. Anna Mapp, Dr. Julie Biteen, and Dr. Pat O'brien. Through our many meetings they have always challenged and encouraged me to go above and beyond in my thinking about the research process. I would also like to thank my fellow Walter lab members who I have had the privilege to get to know, laugh with, and learn from over six wonderful years. Specifically I would like to thank my fellow splicers Mario and Ramya who have made saving the world one spliceosome at a time an enjoyable task.

I am almost certain I would never have survived the struggles and disappointments of graduate school had it not been for my many friends I have been blessed to make while in Ann Arbor: Mario, Marek, Smith, Jim, Cody, Gerwin, Joe, Heidi, John and Liz, and Ryan (Bryan). The memories we have all made together is something I will never forget and something I will always miss. I would also like to thank my friends from college and Ski Trip that have provided a yearly getaway filled with skiing, food, laughter, and male bonding: Mike, Tyler, Elliot, Darryl, Joe, and Bujak.

Last but not least I would like to thank both of my loving families back in Wisconsin and here in Michigan. My parents Diane and Larry and my siblings Daniel and Sami, thank you for continued encouragement, love, and willingness to come visit through these hard years when travel has been hard for me. I will always cherish and appreciate the time we spent together on the Island hunting, fishing, and watching movies. Coming home to cut down the Christmas tree and spend the day hunting are some of my greatest memories and my favorite things to still do today. And I have to thank my second family Barb, Stan, Roger, and Kathy. You have always made me feel like I'm at home while living in Michigan. Thank you for always including me on all your important family events from Thanksgiving to Christmas to the NCAA tournament. And to my amazing girlfriend Tracy: thank you for always supporting me and encouraging me through these very difficult last three years. You have always provided a much needed escape from the stresses that accompany grad school. I am certainly a better person having known you and look forward to our future adventures together.

## Table of Contents

## List of Figures

## List of Tables

## List of Appendices

**Abstract**

Spliceosomes are multi-megadalton RNA-protein complexes that catalyze the removal of introns from pre-messenger RNAs (pre-mRNAs) yielding a continuous protein-coding segment of RNA (mRNA). As a finely tuned process of great complexity and critical importance to the diversification of the proteome, it is thought that up to 50% of all mutations connected to human disease act through disruption of the splicing code.

The structure and conformation of the RNA components of the spliceosome are central to its function. Proper assembly and catalytic activation of the spliceosome require an elaborate sequence of RNA:RNA and RNA:protein rearrangements as well as specific pre-mRNA substrate structures that serve as a scaffold upon which splicing factors and regulators bind to ensure splicing fidelity. Despite 30 years of study, critical questions about the specific structure and conformational rearrangements utilized by pre-mRNA substrates remain unanswered. We have developed a number of biochemical and biophysical approaches that have begun to shed light on pre-mRNA structure during splicing. Using single-molecule immunopurification, we have isolated the activated yeast spliceosome for investigation by single-molecule fluorescence resonance energy transfer (smFRET). Tracking the dynamics of the pre-mRNA during the first catalytic step of splicing revealed a mechanism in which the spliceosome utilizes specific protein cofactors to promote pre-mRNA dynamics in favor of the catalytic conformation. Furthermore, we have dissected the conformational changes at each step of spliceosome assembly and catalysis. Efficient interpretation of the data required development of a single-molecule clustering tool capable of distinguishing FRET states and kinetics. Next, we sought to translate

smFRET trajectories into 3-dimensional RNA structures. Incorporating biochemical footprinting and smFRET data into RNA structure determination, we have begun to model the pre-mRNA structure at each stage of spliceosome assembly. Finally, we have developed a biochemical method that allows for the isolation of *in vivo* assembled spliceosomes and used it to identify a number of new pre-mRNA substrates in yeast. Through the establishment of new biochemical, biophysical, and computational tools for the investigation of splicing, we have finally begun to reveal new molecular mechanisms by which the spliceosome utilizes RNA structure to achieve high efficiency and fidelity.

# CHAPTER 1: Introduction

## 1.1 The RNA Revolution

### 1.1.1 Redefining the Rules Describing RNA Function

DNA makes RNA makes protein[1]. For far too long these have been the words used to describe what is known as the "central dogma of molecular biology" and the words that have been most universally accepted as the way to describe cellular function. Up until the years after the discovery of the double helical structure of DNA by Francis Crick in 1953, very little evidence supported the notion of nucleic acids, particularly RNA, being anything but the carriers of genetic information. Researchers seemed to accept that there are messenger, transfer, and ribosomal RNAs but beyond that, proteins are the workhorse of the cell and responsible for the regulation and execution of all biological processes. Due to this misconception, much of the human genome has often been regarded as "fly over country," regions that are not protein-coding and thus most likely overlooked by the cellular machinery. With these assumptions about the existence and function of RNA came a number of rules by which RNA was governed, the most significant being that only proteins can serve as enzymes[2]. These and other rules about the function of RNA were heavily disputed by those in favor of the RNA world hypothesis, an answer to the origin-of-life problem in which all material originated from RNA precursors and that were responsible for carrying out all chemical reactions required for cellular function. It wasn't until the 1980s that the first renegade non-coding RNAs emerged as a major contradiction

to these long accepted rules. The first of these was the small nuclear RNAs (snRNAs) as potential players in the splicing of introns from pre-messenger RNA[3]. Other abundant classes of non-coding RNA such as small nucleolar RNA (snoRNA) quickly emerged, and achieved great momentum with the discovery of micro RNAs (miRNAs) and the RNA interference pathway[4], revealing a perception of RNA that has evolved from a very primitive role to a complex involvement in nearly every biological process.

The non-coding RNA revolution gained further energy with the completion of the Human Genome Project in 2001[5]. Taking over 13 years and three billion dollars to complete, what resulted was the very first complete sequence of the 3 billion base pairs within the human genome and a plethora of new information that is still being analyzed today[6]. Most interesting was the realization that our genome may contain just over 20,000 protein-coding genes, a mere 2 % of our genome[7,8] and a far smaller number than the 100,000 genes once predicted to code for protein[9]. An obvious question must quickly come to mind: what is the other 98% of the genome responsible for? Recognizing the immense potential modern sequencing techniques have to offer in answering this question, a vast shift in research has begun to focus on the development of methods capable of more cheaply and rapidly investigating the human genome using deep sequencing and has rejuvenated the interest of the scientific community in RNA biology.

### 1.1.2 RNA: the most versatile biomolecule

The discovery of new non-coding RNAs has uncovered roles for RNA that extend from deep within the nucleus where it assists with chromatin remodeling all the way to the extracellular environment where it can act as a signaling molecule[10,11]. We now know that RNA is one of the most structurally and functionally diverse macromolecules, having the ability to serve as a structural scaffold for the assembly of proteins as is the case with the long non-coding RNA

HOTAIR in the ubiquitination pathway[12], interact in a structure- and/or sequence-specific manner with other RNA molecules and individual proteins, and even serve as a small molecule biosensors as is the case with metabolite-responsive ribozymes[13]. In addition, RNA can serve as the mediator of our genetic material from DNA in the nucleus to formation of a functional protein and, in a similar manner, even act as the carrier of genetic material rather than DNA as is the case with many viruses[14]. Furthermore, RNA can serve a catalytic function[15,16] participating in processes requiring RNA cleavage as well as protein biosynthesis.

### 1.1.3 RNA Achieves Function through Secondary Structure

The chemical properties of RNA, specifically the assortment of secondary and tertiary structures, enable it to perform such a diverse range of functions[17]. A single RNA transcript from the same genetic locus can often use a common sequence to carry out multiple roles, or switch between active and inactive forms, simply by altering and mediating its structural features[18,19]. But while the sequence of RNA is the most important determinant of structure, the nucleotide sequence alone is usually not sufficient to specify a unique structural output due to the complex RNA folding pathway. In addition, multiple RNAs with drastically different sequences can form the same or similar structures and thus serve similar functions. Currently, there is a poor understanding of RNA-driven molecular mechanisms in regard to the structure and dynamics of the RNA within biomolecular machines, hindering the development of a proper understanding of how RNA can be modulated and controlled.

## 1.2 The Spliceosome

The spliceosome, a multi-megadalton complex responsible for the majority of RNA splicing, is an ideal example of a macromolecular machine that utilizes nearly all of RNA's many structural and functional features to carry out its purpose. The RNA elements of the spliceosome serve as a

scaffold upon which a multitude of proteins assemble to form the mature snRNP components. These snRNPs utilize sequence specificity to recognize the pre-mRNA substrate splice sites and carry out both chemical steps of splicing through what is thought to be an RNA catalyzed mechanism[20]. A thorough understanding of how the spliceosome utilizes and modulates RNA structure to ensure catalytic efficiency is central to understanding its role in gene expression, regulation, and disease. Additionally, mechanisms utilized by the spliceosome are common among RNA-based machines and could thus provide us with a more complete understanding of all cellular processes.

### 1.1.1 Introns: the 'Where's Waldo' of our genetic code

Traditionally it was thought that mRNA was simply a direct copy of the DNA gene encoding a particular protein, and that one gene denotes only one protein output. It was not until the 1970s that Philip Sharp and Richard Roberts, utilizing RNA-DNA hybridization and electron microscopy, discovered that large segments of the adenovirus 2 genome remain un-base paired when hybridized with the purified mRNA product[21,22]. This revealed that transcribed messenger RNAs are not necessarily a copy of continuous segments of a DNA gene, but rather pieced-together bits of coding region with large gaps in the sequence. The removed fragments of genomic code were later termed introns, with the coding segments termed exons.

Today, it has become increasingly clear that the presence of introns is crucial to the function of eukaryotic organisms, as removal of all introns from yeast has extremely deleterious effects[23]. Upon completion of the human genome project, we now know that nearly 95% of the human transcriptome consists of introns[24] and that approximately 92-95% of multi-exon genes in humans are alternatively spliced[25,26]. Because alternative splicing allows for the expression of multiple protein isoforms from a single RNA transcript, the human genome can be compacted

into a much smaller sequence than would be expected given our much increased complexity compared with lower eukaryotic organisms. In addition, alternative splicing provides multiple points of regulation over gene expression.

Aside from self-splicing group II introns, all introns are recognized and spliced by the spliceosome in what is thought to be primarily a co-transcriptional manner[27,28]. Spliceosomes are assigned the task of very rapidly finding the short, often divergent, RNA sequences designating important splice sites within an RNA transcript that can be up to tens of thousands of nucleotides long. The spliceosome must then rapidly assemble and remove only specific introns in order to achieve cell- and tissue-specific expression of certain protein isoforms. Interestingly, and perhaps not surprisingly, these introns are not simply discarded segments of RNA junk without a future purpose or function[29]. Often introns are debranched and processed into other RNAs such as small nucleolar RNA (snoRNA), microRNAs, and long noncoding RNAs where they serve functions in translation, transcription, and gene regulation. Given their presence in eukaryotic genomes and their persistence through evolution, it is clear that introns and the splicing process serve an important role in cellular biogenesis. Having a thorough understanding of intron structure, dynamics, and sequence elements that enhance or inhibit the splicing process is crucial to fully understanding how large RNA machines like the spliceosome function and how it can be fine-tuned for therapeutic purposes.

### 1.1.2 The Complexity of Spliceosome Assembly and Catalysis

The budding yeast *Saccharomyces cerevisiae* spliceosome is composed of nearly 80 proteins and 5 small nuclear RNAs (snRNAs) referred to as U1, U2, U4, U5 and U6. The spliceosome recognizes the pre-mRNA substrate utilizing three conserved regions, namely the 5' splice site (5'SS), branch site (BS), and the 3' splice site (3'SS) (**Figure 1.1**). The first chemical step of


**Figure 1.1 The Conserved Yeast pre-mRNA Substrate**
The yeast pre-mRNA 5' splice site (5'SS), branch site (BS), 3' splice site (3'SS) are all highly conserved regions required for splicing. The branch point adenosine is underlined.

splicing involves the nucleophilic attack of the 5'SS by the 2' hydroxyl (2'OH) of the BS adenosine residue, yielding a free 5' exon and an intron lariat structure containing a 2'-5' phosphodiester bond at the branch site adenosine immediately upstream of the 3' exon. The spliceosome is then rearranged to allow the free 5' exon to undergo a nucleophilic attack on the 3'SS in the second catalytic step, resulting in release of the ligated exons (mRNA) and the intron lariat (**Figure 1.2)**.

In contrast to other macromolecular machines such as the ribosome, the spliceosome does not have a pre-formed catalytic core. Rather, each of the five small nuclear ribonucleoprotein complexes (snRNPs, denoted U1, U2, U4, U5, and U6), which are themselves composed of one of the five snRNAs and several associated proteins, assemble upon a single pre-mRNA substrate in a stepwise fashion, carry out both chemical steps of splicing, and then disassemble to carry out further rounds of splicing on other pre-mRNAs. Such a stepwise assembly process allows for tight regulation of the splicing process by providing multiple checkpoints before, during, and after both steps of splicing. Assembly begins with the recognition of the 5'SS by the U1 snRNP complex in an ATP-independent fashion through specific base pairing of the U1 snRNA with the 5'SS[30] (**Figure 1.3**), followed by the sequence specific recognition of the BS by the branchpoint binding protein (BBP)-Mud2 dimer (Commitment Complex 2, CC2). In the first ATP-dependent assembly step, the U2 snRNP is loaded onto the BS with the assistance of the Sub2 and Prp5 ATPases (A complex). The U4, U5, and U6 snRNPs bind to this A complex structure as a preformed complex known as the tri-snRNP to form the B complex. Large RNA-RNA and RNA-protein rearrangements occur at this point such that the RNA-RNA base pairing between the 5'SS and U1 snRNA is disrupted, resulting in destabilization and removal of the U1 and U4 snRNPs and binding of the U6 snRNP to the 5'SS. Assembly proceeds with the joining of the

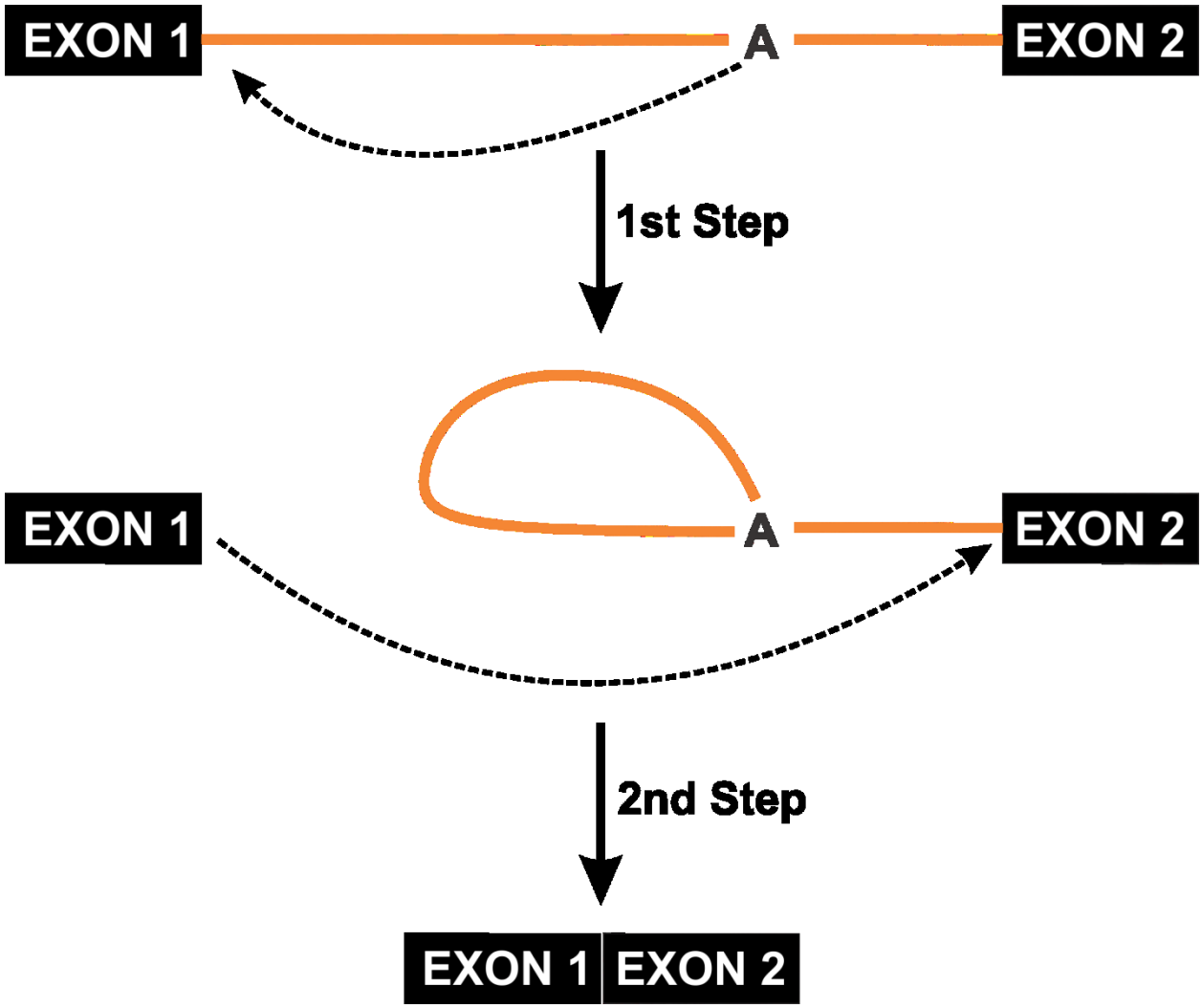



**Figure 1.2 The Two Chemical Steps of Splicing**
Pre-mRNA splicing takes place in two subsequent catalytic steps. The first step involves the nucleophilic attack of the 2'OH of the BS adenosine on the 5'SS releasing a free 5' exon (exon 1) and an intron lariat intermediate. The second step proceeds with the attack of the free 3'OH of the 5' exon on the 3'SS, releasing ligated mRNA

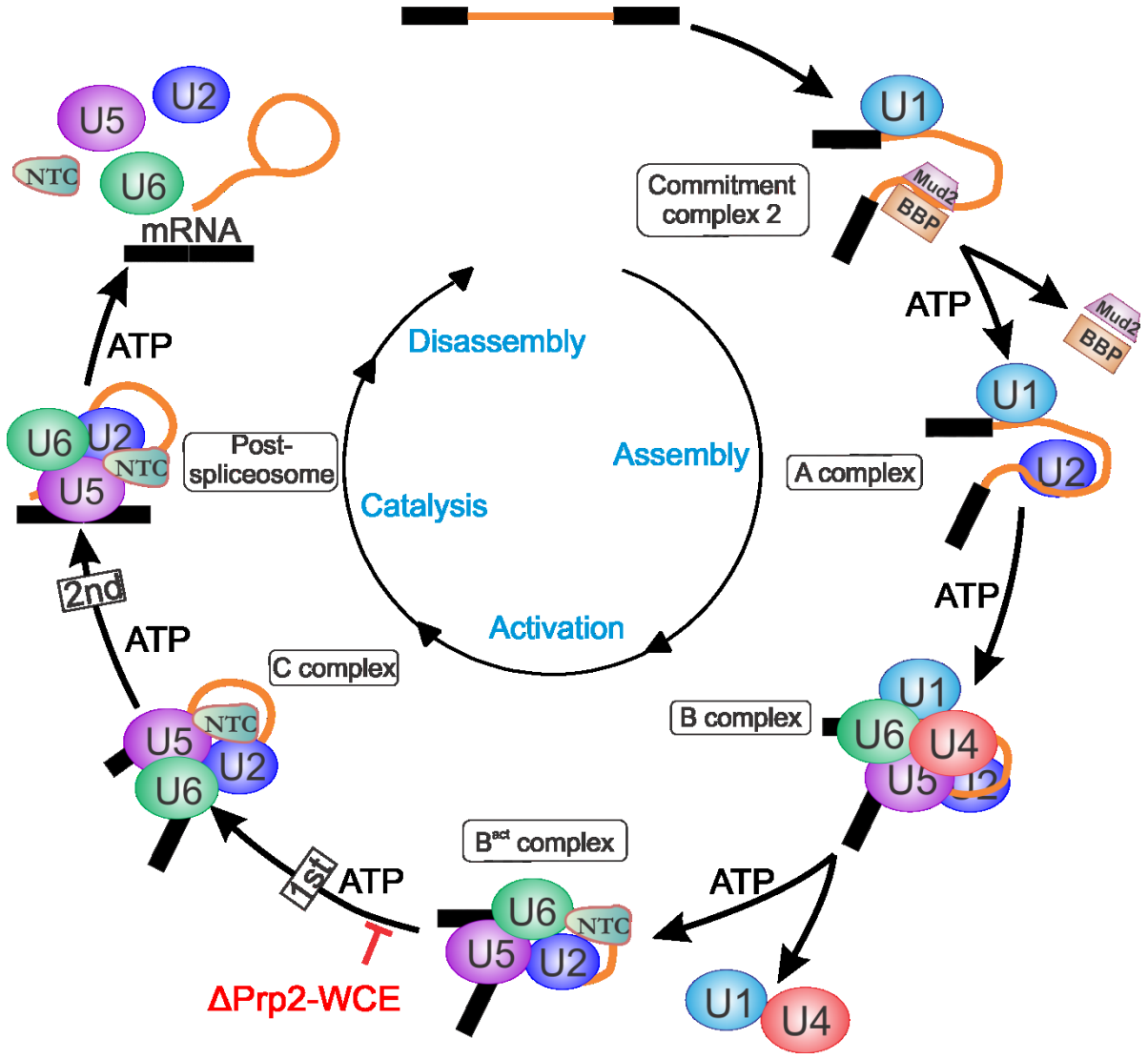



**Figure 1.3 The Canonical Spliceosome Assembly Pathway**
Schematic representation depicting the stepwise assembly of the spliceosomal subunits on a pre-mRNA substrate. Assembly, activation, catalysis, and disassembly steps are highlighted as are ATP-dependent steps in assembly. The heat-sensitive Prp2 mutation used in a majority of this thesis is indicated, resulting in accumulation of the yeast $B^{act}$ complex[36,37,59,65].

snRNA-free NineTeen Complex (NTC) yielding the fully activated spliceosome ($B^{act}$). Although entirely assembled and activated, catalysis requires the addition of Prp2 in order to carry out a significant reorganization of the catalytic core. Together with its cofactor Spp2, Prp2 utilizes ATP in order to form the B* complex which is now capable of carrying out the first step of splicing. Association of Cwc25 then stabilizes the C complex, a highly reactive conformation of the spliceosome capable of efficient first step catalysis, yielding the free 5' exon and lariat intermediate. The catalytic core is then reorganized such that the 5'SS and 3'SS are placed into close proximity, allowing the spliceosome to carry out efficient ligation of the free 5' exon to the 3' exon in the second chemical step of splicing. Lastly, Prp43 and Prp22 are utilized to completely disassemble the spliceosome, resulting in release of the ligated mRNA and intron lariat, as well as recycling of the spliceosomal components.

### 1.1.3 The role of ATP and Dynamics in Splicing

Although both chemical steps of splicing are isoenergetic, ATP is consumed by at least 8 known RNA-dependent ATPases responsible for the complex RNA-RNA and RNA-protein rearrangements crucial to substrate identification and accurate spliceosome assembly (**Figure 1.4**). These ATPases are most similar to the DExD/H box family of helicases with targets and mechanisms of action that are often still unknown.

The earliest ATP-dependent step is the Prp5- and Sub2-mediated association of the U2 snRNP with the BS through base pairing between the U2 snRNA and BS to form the A complex. Such an association requires the removal of the BBP-Mud2 heterodimer from the BS by Sub2. Interestingly, Sub2 has been shown to first help stabilize this interaction in an ATP-independent manner, but then ultimately uses ATP hydrolysis to remove Mud2 and BBP prior to U2 assembly[31,32]. While Sub2 frees up the BS for U2 snRNP binding, Prp5 is responsible for

stabilizing the U2 snRNA interaction with the BS and does so in a multidimensional fashion. The U2 snRNA possesses two conserved regions that must adopt specific conformations to stably bind the pre-mRNA. The first is a stem loop known to transition between a Cus2 stabilized, inactive conformation (IIc) and a Cus2-free, binding-active conformation (IIa). In its first proposed ATP-dependent role, Prp5 hydrolyzes ATP to bring about removal of Cus2, formation of stem loop IIa, and stable binding of U2 to the BS. Given a recent study that showed evidence for Prp5 presence throughout the splicing cycle[33], Prp5 may even perform further rounds of stem loop IIa-IIc interconversion to promote proper assembly and catalysis throughout the splicing process[19]. The second is the evolutionarily conserved branchpoint-interacting stem loop (BSL), a U2 RNA structural element that displays the BS interacting region of U2 through formation of base pairing between the adjacent regions[34]. The BSL is thought to play an important role with Prp5 in splicing assembly during U2 snRNP recruitment. Prp5 again utilizes ATP hydrolysis to disrupt the BSL and allow complete recognition of the intron by the U2 snRNA. Prp5 has been proposed to be involved in proofreading the substrate at this stage as hyperstabilization of the BSL slows the ATP-dependent unwinding of the BSL, allowing for suboptimal intron BS sequences to be utilized for splicing[34,35].

Upon incorporation of the U4/U5/U6 tri-snRNP, an extensive network of RNA-RNA and RNA-protein interactions must be broken and then reformed such that the U1 snRNP becomes displaced from the 5'SS allowing for the U6 snRNP to take its place, a process that utilizes the ATPase activity of Prp28 and Brr2 (**Figure 1.4**)[36]. Prp28 hydrolyzes ATP in order to remove or at least destabilize the U1 snRNP from the 5'SS. The U6 snRNA joins the spliceosome tightly base paired to the U4 snRNA, which serves more of a shuttling role, preventing U6 snRNA from prematurely recognizing and binding the 5'SS. Brr2 is the helicase responsible for unwinding the

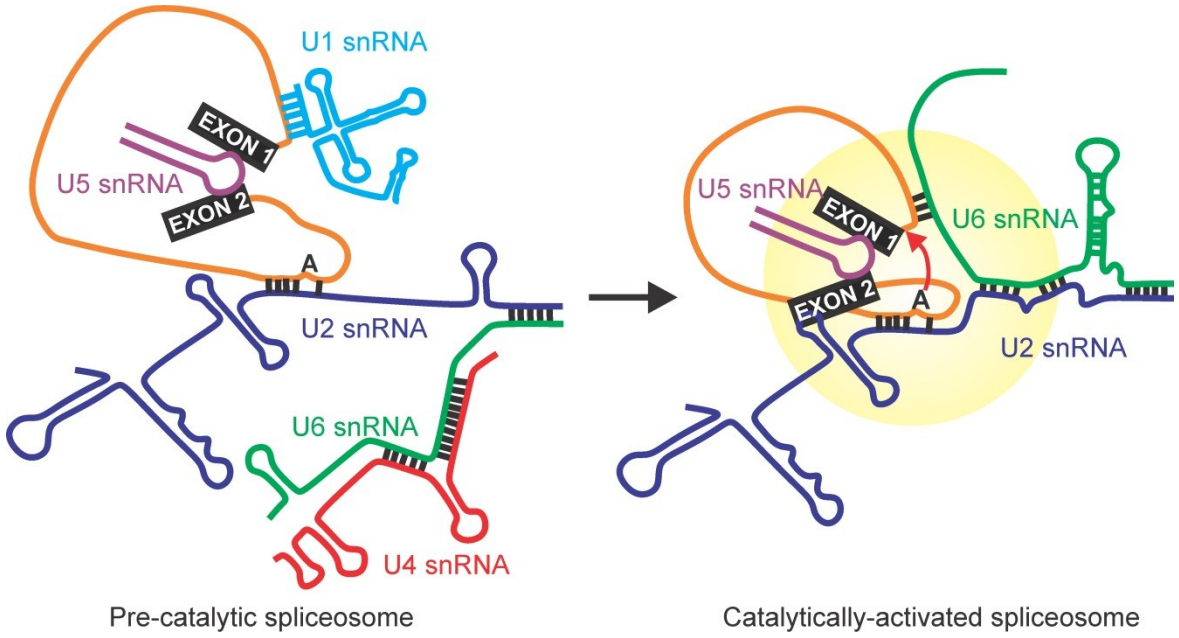



**Figure 1.4 RNA-RNA Rearrangements during Spliceosome Activation**
Extensive RNA-RNA rearrangements are required with the snRNA components of the spliceosome as well as with the pre-mRNA substrate to allow catalytic activation of the spliceosome. Figure adapted from Lührmann *et al.*[36]

U4/U6 duplex, allowing U6 to bind the now accessible 5'SS and the U1 and U4 snRNPs to completely dissociate from the spliceosome, resulting in formation of the activated conformation of the spliceosome ($B^{act}$).

Prp2 was first identified as one of several temperature-sensitive mutations in yeast that resulted in a splicing defect[37]. As their name implies, pre-mRNA precursor mutants, or prp mutants, accumulate precursor or splicing intermediates, suggesting that the specific mutation results in the protein being incapable of carrying out a specific function and thus certain step(s) of splicing. Although this particular prp2-1 allele was found to result in accumulation of pre-mRNA upon shift to the non-permissive temperature, it was later found that Prp2 mutant alleles completely assemble into catalytically active spliceosomes ($B^{act}$) whose activity can be restored upon addition of recombinant proteins Prp2, Spp2, Cwc25, and ATP[38]. The ATPase Prp2 required at this step was found to cause a dramatic shift in spliceosome sedimentation coefficient upon ATP hydrolysis, suggesting either a significant loss of protein or a dramatic change in spliceosome conformation[39]. Proteomic and fluorescence microscopy analysis later revealed there is indeed no significant loss of protein, suggesting that the change in sedimentation coefficient is more than likely due to a change in conformation[40]. This new conformation, termed the B* complex, is capable of carrying out low levels of first step splicing, indicating that the action of Prp2 allows the spliceosome to experience conformations suitable for splicing. Only upon addition of Cwc25 does this first-step-capable conformation become stabilized and the first step of splicing can efficiently take place. The mechanism by which Prp2, Spp2, and Cwc25 carry out their function is discussed in **Chapter 2**. Following the first step of splicing, the spliceosome core must be reorganized to allow attack of the 5' exon on the 3'SS for exon ligation. This begins with ATP hydrolysis by Prp16 to remove Cwc25 from the BS so that the

second step factors can bind[41]. Like Prp2, Prp16 is thought to allow for a significant amount of proofreading at this stage, a discovery made when Prp16 mutants allowed mutated substrates to proceed through splicing in what was later termed the kinetic proofreading model[42]. It is hypothesized that the mutant Prp16 allows a suboptimal substrate to spend ample time in a conformation that promotes first step splicing with Prp2, Spp2, and Cwc25 present, consequently allowing splicing to occur even with a mutation in the BS. Conversely, normal Prp16 shifts the catalytic core towards the second step conformation upon ATP hydrolysis before a mutated substrate has the time to be spliced so that instead it is rejected.

Further spliceosomal proofreading is provided by Prp22, which is thought to operate using a similar mechanism but detect mutations that affect exon ligation[43]. Prp22 hydrolyzes ATP to promote mRNA release following exon ligation. When an optimal 3'SS is detected, exon ligation will occur faster than Prp22 can hydrolyze ATP and, as a result, Prp22's ATPase activity will release ligated mRNA from the spliceosome. However, the presence of a mutated 3'SS in the catalytic core will stall splicing immediately prior to the second step of splicing. ATP hydrolysis by Prp22 now occurs faster than exon ligation, resulting in the release of unspliced RNA from the spliceosome instead of the ligated mRNA. Given Prp22's high *in vitro* helicase activity, it is thought that this release occurs through unwinding of the mRNA from the U5 snRNA. Lastly, Prp43 utilizes ATP hydrolysis to completely dissociate the spliceosome allowing for further rounds of assembly and catalysis.

## 1.2 Visualizing RNA Structure and Dynamics

### 1.2.1 Single molecule FRET

Any technique used to study a large, dynamic macromolecular complex such as the spliceosome must be sufficiently sensitive to detect low concentrations of sample, sufficiently specific to

address a particular location of interest amid a large background of other protein and RNA components, and sufficiently information-rich to allow rigorous testing of mechanistic hypotheses. One technique that meets these requirements is single-molecule fluorescence (or Förster) resonance energy transfer (smFRET)[44-47]. In a FRET experiment, a sample is labeled with a pair of fluorophores, chosen so that the emission spectrum of one (the "donor fluorophore") overlaps with the absorption spectrum of the other (the "acceptor fluorophore"). When the donor is excited, a dipole-dipole interaction between it and the acceptor permits the transfer of energy between the two, with the efficiency of this process depending on the distance and relative orientation between the two fluorophores, the fluorescence quantum yield of the donor, and the extent of spectral overlap between the donor's emission and the acceptor's absorption (**Figure 1.5a**). This distance dependence, which has a sensitivity range of ~10-100 Ångstroms, makes FRET a valuable technique for probing the conformations and conformational dynamics of biological macromolecules by monitoring changes in FRET efficiency[46-48]. In single-molecule FRET, the molecule of interest is immobilized sparsely on a microscope slide so that the donor and acceptor fluorescence intensities, and thereby the FRET efficiency, can be measured for individual molecules (**Figure 1.5b,c**)[45]. This is valuable because complex biological macromolecules often exist in multiple different conformations, and single-molecule FRET allows these conformations and their transitions to be observed, rather than reporting an ensemble average, which loses most of the information on transition kinetics, transient intermediates, and rare conformational states[45,49].

In recent years, smFRET has been applied to a number of protein-RNA complexes, including but not limited to the bacterial ribosome[50-52], the yeast spliceosome[53-55], and human telomerase[56,57]. In most smFRET work, purified nucleic acid and protein components have been

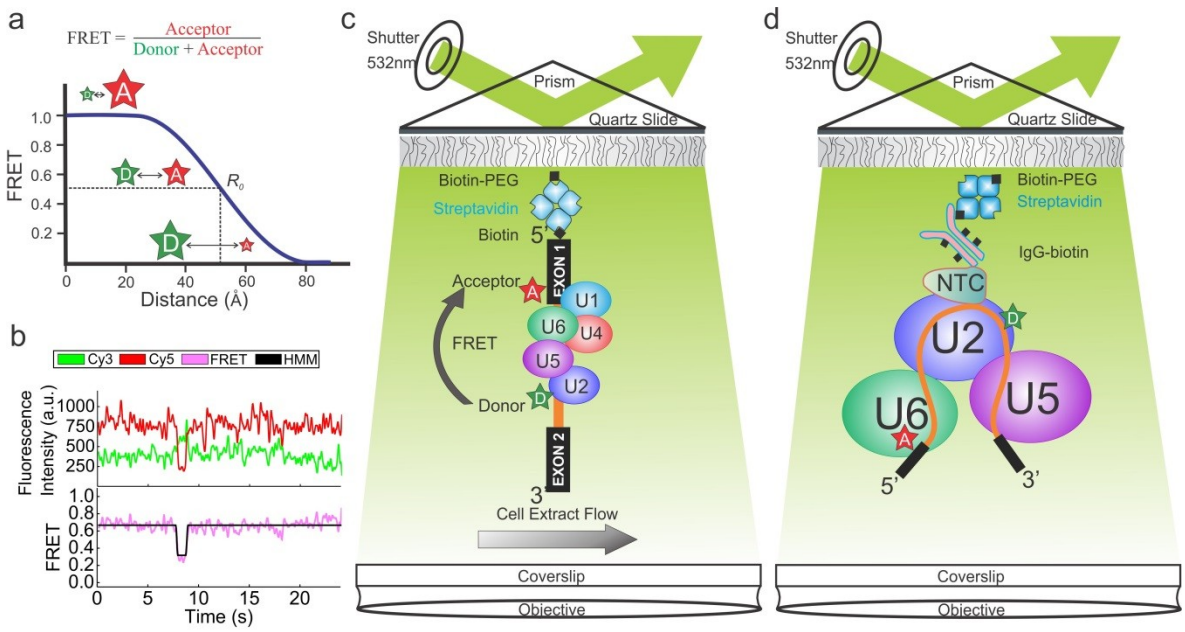



**Figure 1.5 Using single molecule FRET to observe RNA dynamics**
(a) Distance-dependent changes in FRET utilizing donor (D) and acceptor (A) fluorescent dyes. (b) Typical smFRET trajectory depicting the anti-correlated nature of the donor (Cy3) and acceptor (Cy5) fluorophores. The corresponding FRET trajectory is fit to a Hidden Markov Model (HMM) using vbFRET software. (c) Monitoring time- and ATP-dependent changes in pre-mRNA conformation as the spliceosome assembles and carries out both steps of splicing on an immobilized substrate. (d) SiMPull-FRET immobilization of the $B^{act}$ complex

immobilized on slides through biotin-streptavidin linkages. Even the comparably small yeast spliceosome contains so many different components (five different snRNAs and ~40 different proteins, depending on the stage of splicing)[58,59] that it is not practical to purify every protein and RNA component in the spliceosome and reconstitute splicing "from scratch". Conversely, when working in cell extract, splicing can be stalled at many different stages of the splicing cycle using, for example, genetic manipulations that are readily available in yeast, leading to accumulation of certain intermediate complexes that can then be isolated and subjected to biochemical analysis. The gap between these two areas of inquiry (single molecule observation of purified components and biochemical analysis of complexes isolated from cell extracts) was bridged by the technique of single-molecule pull-down (SiMPull), which was first demonstrated in 2011[60,61]. In this approach, a streptavidin-coated slide is incubated with a biotinylated antibody, and extract is prepared from cells bearing a matching epitope, for example, a TAP or FLAG tag, on a protein of interest. This extract is incubated on the slide, allowing a particular complex to be "pulled down" from the extract onto the slide through the interaction between the antibody and the epitope. An extension of this approach termed single-molecule pull-down FRET (SiMPull-FRET) allows complexes to be studied via smFRET that otherwise may be difficult to purify, immobilize and/or reconstitute, and offers the potential for them to be studied in cell extract (**Figure 1.5d**)[55]. This work will be thoroughly discussed in **Chapter 2.**

A current challenge in the single molecule field is the proper and thorough analysis of complex datasets containing a large number of FRET states from several conditions with intricate kinetics. The classical analysis methods utilizing histogram and TODP analysis work great for understanding the average FRET behaviors within a population but fall short when subpopulations of molecules show distinct kinetics and more than two reversibly interchanging

states. An alternative analysis technique termed single molecule cluster analysis (SiMCAn)[62] that allows for the complete interpretation of complex smFRET data will be discussed in **Chapter 3**.

There is also a great interest in the smFRET and RNA fields to be able to project the FRET states and trajectories found for a particular RNA into predicted 2D and 3D structural pathways in the RNA folding landscape. Current theoretical research in RNA folding kinetics lacks support from experimental data. In addition, structure prediction software that is available can only incorporate biochemical footprinting data to modify predictions. The ability to translate a time series of FRET states and transitions into predicted changes in RNA structure would help give a more real picture of the structure of Ubc4 in each of the assembly and catalytic stages of splicing based solely on gathered smFRET data. A detailed description of a new method we have developed is provided in **Chapter 4**.

1.2.2 Multiplexing through deep sequencing-coupled footprinting

When measuring by smFRET the conformational dynamics of the Ubc4 pre-mRNA[53,62], it was noticed that the RNA dynamically folds into long-lived high-FRET states even in splicing buffer alone, suggesting that the fluorophore-labeled exons are readily positioned much closer than expected from their linear sequence separation imposed by the 95-nt long intron. This observation is in accord with recently accumulating evidence that supports the hypothesis that intron secondary structure has a functional role in splicing. Unfortunately, smFRET measurements of a pre-mRNA's structure only report on the relative distance between segments of a single RNA substrate, making the verification of such a hypothesis very difficult and time-consuming. Recently, a significant amount of effort has been put forth to develop cheaper and more high-throughput approaches that allow for the investigation of RNA structure by combining RNA structure probing techniques (SHAPE, DMS, enzymatic probing) with

massively parallel sequencing to read out the results of a single RNA as well as entire transcriptomes. One such approach, selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP)[63], utilizes an RNA modifying reagent to produce 2'-modified RNA within accessible, single-stranded regions of the RNA. These modifications are detected during deep sequencing and used as restraints in RNA structure prediction algorithms. A detailed description of first steps toward the use of SHAPE-MaP to investigate the intron structure hypothesis is provided in **Chapter 5**.

# CHAPTER 2: Biased Brownian Ratcheting leads to pre-mRNA Remodeling and Capture prior to the First-Step of Splicing[1]

## 2.1 Introduction

Introns are removed by the spliceosome, a large ribonucleoprotein (RNP) complex, in a two-step transesterification process. In the first step, the 2'OH of the branchpoint adenosine (BP) attacks the phosphodiester bond at the 5' splice site (5'SS), releasing the 5' exon and creating the branched, lariat structure; in the second step, the 3' hydroxyl of this exon attacks the phosphodiester bond at the 3' splice site (3'SS), releasing the lariat intron and creating the spliced mRNA product[64]. The most conspicuous feature of this enzyme is that it lacks a preformed catalytic core, which is created in a stepwise fashion, beginning with the assembly of the U1 and U2 small nuclear RNPs (snRNPs) at the 5'SS and BP, respectively, to form the pre-spliceosome (A complex)[59]. The U4–U6.U5 tri-snRNP then binds to create the mature spliceosome (B complex)[59]. Notably, however, U1 and U4 snRNPs must be removed before catalysis, creating first the activated B ($B^{act}$) complex and then, after additional rearrangements, the catalytically active B* complex. The resulting post-first-step (C) complex then undergoes further remodeling required for the second step of splicing and the formation of mature mRNA[59].

[1] Adapted from Krishnana, R., Blanco, M., Kahlscheuer, M., Abelson, J., Guthrie, C., Walter, N.G. "Biased Brownian ratcheting leads to pre-mRNA remodeling and capture prior to first-step splicing. *Nat. Struct. Mol. Biol.* **20,** 1450-1457 (2013). Matthew Kahlscheuer performed all protein expression, purification, and fluorescent labeling, biochemical experimentation and validation of SiMPull-FRET, and smFRET experiments and data analysis involving labeled Cwc25. Ramya Krishnan performed all other smFRET experiments and data analysis. Mario Blanco performed biochemical experimentation and developed the MATLAB scripts used in this work.

The highly dynamic process of spliceosome assembly and catalysis is guided by a set of RNA-dependent ATPases of the DExD/H-box helicase family that collectively function to insure the fidelity of splicing[65]. A major experimental challenge has been to understand the precise conformational rearrangements of RNA and protein that accompany each ATP-dependent step. DExD/H-box helicase Prp2 is required for the first chemical step of splicing, and recent proteomic analyses of the $B^{act}$, B* and C complexes revealed that its action results in the destabilization of the U2 snRNP-associated proteins SF3a and SF3b[38-40,66]. An attractive hypothesis is that SF3b sequesters the BP adenosine[67] to prevent a premature attack on the 5'SS, in which case the ATP-dependent action of Prp2, together with its cofactor Spp2, would be required to initiate catalysis. In a biochemical *tour de force*, successful reconstitution of both steps of splicing with the addition of recombinantly expressed proteins to immunopurified splicing complexes has been demonstrated[38,68,69]. In particular, first-step chemistry could be achieved with the addition of ATP, Prp2, Spp2 and Cwc25[38].

We set out to investigate the roles Prp2, Spp2, and Cwc25 have in activating the spliceosome for the first step of splicing, developing an approach that couples the purification of specific splicing complexes with single-molecule fluorescence resonance energy transfer (FRET). We used the resulting single-molecule pulldown FRET (SiMPull[70]-FRET) technique to analyze a functional $B^{act}$ complex assembled on a pre-mRNA with fluorophores near the scissile bonds. When this complex is assembled in an extract with a temperature-sensitive allele of Prp2 (*prp2-1*)[37], the $B^{act}$ spliceosome is stalled and can catalyze the first step of splicing only through formation of the B* complex, upon addition of ATP, Prp2 and Spp2, and then the C complex, upon further addition of Cwc25. Using SiMPull-FRET, we show that ATP-dependent action of Prp2 and its cofactor Spp2 unlocks reversible switching of the intron between splicing-active and

splicing–inactive conformations. Cwc25 then rectifies this thermal Brownian ratcheting by stabilizing the conformation in which the 5'SS and BP are in close proximity, driving the equilibrium toward catalysis.

## 2.2 Materials and Methods

### 2.2.1 Affinity purification of the $B^{act}$ complex

Extracts were prepared from a *prp2-1 cef1-TAP* yeast strain (ATCC 201388: *MAT***a** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*[71]), heated at 37 °C for 40 min to inactivate Prp2 and stall the spliceosome at the $B^{act}$ complex. In a final volume of 135 µl, 40% (v/v) of this heat treated extract was incubated with ~50 pmoles FRET labeled Ubc4 pre-mRNA[53] in the presence of 2 mM ATP in splicing buffer (8 mM HEPES-KOH, pH 7.0, 2 mM $MgCl_2$, 0.08 mM EDTA, 60 mM $K_i(PO_4)$, 20 mM KCl, 8% (v/v) glycerol, 3% (w/v) PEG, 0.5 mM DTT) and incubated at 23 °C for 35 min. For biochemical experiments, streptavidin-coated magnetic beads (Dynabeads® MyOne™ Streptavidin C1, Invitrogen) were handled as per the manufacturer's recommendation. For each splicing reaction, 200 µl of the suspended beads were equilibrated in 200 µl of T50 buffer (50 mM Tris-HCl, pH 7.5, 50 mM NaCl). An equal volume of 0.5 mg/ml biotin-IgG (ZyMAX™ Rabbit Anti-Mouse IgG (H+L) - BT (ZyMAX™ Grade)) in T50 was added and incubated in a tube rotator at 23°C for 30 min. The beads were then pulled down using a magnet and the supernatant was discarded. To block any streptavidin not bound by biotin-IgG, the beads were incubated with excess free biotin at 1.5 mg/ml in T50 buffer in a tube rotator at 23°C for 20 min. After equilibration in splicing buffer, the independently assembled splicing reaction were added and incubated in a tube rotator for 30 min at 23°C to allow the protein A of the Cef1-TAP tag in the spliceosome complex to bind the biotin-IgG. Upon removal of the supernatant, the beads were further washed three times with buffer A (20 mM HEPES-KOH, pH 7.9, 120 mM

KCl, 0.01% NP40, 1.5 mM MgCl2, 5% (v/v) glycerol) and once with splicing buffer to further purify the $B^{act}$ complex. The reactions are scaled up for reconstitution reactions pursued in parallel and split at this step. Prp2, Spp2 and Cwc25 were added at 90-120 nM final concentration in splicing buffer in the presence or absence of 2 mM ATP or AMPPNP or UTP and incubated in the tube rotator for 30-40 min for various levels of reconstitution. RNA was isolated and products of splicing were analyzed on a denaturing, 7 M urea, 15% polyacrylamide gel and scanned on a Typhoon variable mode imager (GE Healthcare). Normalized product for Ubc4 was calculated by taking the amount of free 5' exon and dividing that by the total amount of pre-mRNA and free 5' exon. Normalized product for Actin was calculated by taking the amount of lariat intermediate and dividing that by the total amount of pre-mRNA and lariat intermediate.

2.2.2 Cloning, expression, and purification of splicing factor proteins

The full-length PRP2 gene was PCR-amplified and ligated into plasmid pRSETa (Invitrogen) with a C-terminal hexahistidine tag. The N-terminally truncated form of SPP2 (coding for amino acids 37-185) containing a C-terminal hexahistidine tag, and the full-length Cwc25 gene were obtained from Reinhard Lührmann (Max Planck Institute for Biophysical Chemistry, Germany). Cwc25 was subcloned into a pRSETa plasmid containing a single cysteine residue and hexahistidine tag at the C-terminus. The constructs were then transformed into Escherichia coli strain Rossetta II (Novagen). Cultures were grown in 2-4 L of TB medium and induced with 125 μM IPTG. Cultures were then incubated at 20 ºC for 18 h. Cells were harvested, washed, and the pellets stored at -80 ºC.

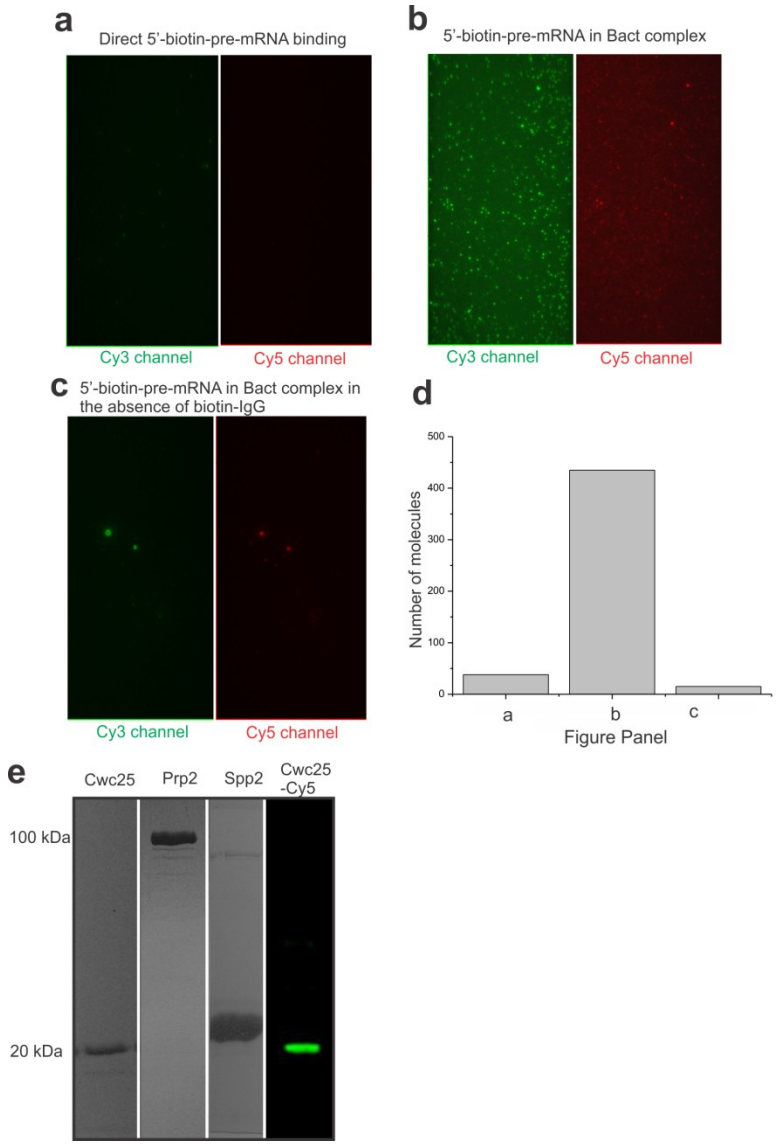



**Figure 2.1 Binding specificity of $B^{act}$ complex and purified proteins used for reconstitution**

(a) Field of view showing direct binding of the 5' biotinylated pre-mRNA to the streptavidin on the slide surface saturated with biotin-IgG and free biotin. (b) Field of view showing the binding of the immunopurified $B^{act}$ spliceosome (with Cef1-TAP) to the streptavidin on the slide surface saturated with free biotin and biotin-IgG. (**c**) Field of view showing the binding of the immunopurified $B^{act}$ spliceosome (with Cef1-TAP) to the streptavidin on the slide surface saturated with free biotin in the absence of IgG-biotin. Left and right panels are the Cy3 and Cy5 channels, respectively. A 532 nm and 635 nm laser was used for excitation under all conditions. (d) Quantification of number of molecules under conditions a-c. (e) Protein expression and purification confirmed by SDS-PAGE analysis. Histidine-tagged Cwc25, Prp2 and Spp2 are shown in lanes 1, 2 and 3, respectively. Cy5-fluorophore labeled single-cysteine Cwc25 is shown in lane 4.

| | |
|---|---|
| UBC4Wildtype (WT) | 5'-GAACUAAGUGAUC**(5-N-U)**AGAAAG*GUAUGUCUAAAGU UAUGGCCACGUUUCAAAUGCGUGCUUUUUUUUUAAAACU UAUGCUCUUAUUUACU**A**ACAAAA(5-N-U)CAACAUGCUAUUG AACUA**G**A**G**AUCCACCUACUUCAUGUU-3'* |
| UBC4 3'Splice Site mutant (3'SS) | 5'-GAACUAAGUGAUC**(5-N-U)**AGAAAG*GUAUGUCUAAAGU UAUGGCCACGUUUCAAAUGCGUGCUUUUUUUUUAAAACU UAUGCUCUUAUUUACU**A**ACAAAA(5-N-U)CAACAUGCUAUUG AACUA**C**A**C**AUCCACCUACUUCAUGUU-3'* |
| DNA splint (dSplint) | 5'-GTTGATTTTGTTAGTAAATAAG(SP9)GTTTTAAAAAAAAAGCACGC-3' |

**Table 2.1 Sequence information of oligonucleotides used in this study**
The Ubc4 intron is italicized, and the allyl-amine modified uridines are denoted as (5-N-U). The red and green colors represent positioning of the Cy5 and Cy3 fluorophores, respectively. In the 3'SS mutant, the two bold and underlined cytosines replace guanines in the wildtype 3' splice site. The bold and underlined A is the BP adenosine. dSplint is the DNA splint used for templated ligation to synthesize the pre-mRNA as described[53]. Sp9 denotes a 9-carbon linker.

Purification of Cwc25-His and Spp2-His was performed as described[38]. Protein purity was confirmed by 16% SDS-PAGE (**Figure 2.1**) and proteins were either first fluorescently labeled or directly aliquoted, flash frozen in liquid nitrogen, and stored at -80 ºC. Protein concentrations were determined by Bradford assay and measurement at $A_{280}$[72]. His-tagged Prp2 obtained from E.*coli* cell lysate was purified as described. Protein purity was confirmed by 10% SDS-PAGE (**Figure 2.1**) and the final product aliquoted, flash frozen in liquid nitrogen, and stored at -80 ºC. Protein concentrations were determined by Bradford assay and measurement at $A_{280}$. RNA sequences used in the study are reported in **Table 2.1**.

2.2.3 Single-molecule FRET of purified spliceosomal complexes

For the single molecule FRET experiments on affinity-purified complexes, we prepared slides using previously published procedures[73]. In short, the surface of a quartz slide was amino functionalized, PEGylated and reacted with 0.2 mg/ml streptavidin in T50 buffer for 15 min at 23°C. 100 µl of 0.5 mg/ml biotin-IgG in T50 was flowed onto the slide and incubated for 20 min, followed by free biotin at 1.5 mg/ml in T50 buffer for 15 min. $B^{act}$ spliceosomal complexes were assembled and stalled as described above by incubation of FRET labeled Ubc4 pre-mRNA with heat treated *prp2-1 cef1-TAP* yeast splicing extract in splicing buffer supplemented with 2 mM ATP and an oxygen scavenger system (OSS) composed of protocatechuate dioxygenase, protocatechuate and Trolox[53]. These complexes were then flowed onto the slide surface and incubated for 15-20 min to allow the Cef-1-TAP on the spliceosome to bind biotin-IgG. The slide surface was washed rigorously and reconstituted (in the presence of OSS) as described for the biochemical purification and incubated for 10-40 min before acquiring data. A home-built prism-based TIRF microscope was used to collect data as described[53,74,75]. To obtain FRET data, we directly excited the Cy3 donor near the BP adenosine with a 532-nm laser, and we recorded

emission by Cy3 and Cy5 fluorophores at 100-ms time resolution using an intensified CCD camera (I-Pentamax, Princeton Instruments).

2.2.4 Fluorescent labeling of Cwc25 and distance estimation from FRET

The single-cysteine mutant of Cwc25 was labeled with Cy5-maleimide (GE Healthcare). Labeling was performed using 0.150 μmol of Cwc25 in storage buffer and 0.5 mg of dye containing 10 μM reducing agent Tris(2-carboxyethyl)phosphine (TCEP) (Sigma). Reactions were incubated at 23°C for 1 h followed by overnight at 4 ºC. Free dye was removed by re-purification of protein on a $Ni^{2+}$ column and dialysis back into storage buffer. The degree of labeling was determined using GE Healthcare's protocol and was found to be 70%. Protein functionality was confirmed using an ensemble pull down assay as described above. The fluorophore distance, R, and the apparent FRET efficiency, $E_{app}$, were calculated as described[74,76,77] from the equations $E_{app} = c[1 + (R/R_0)^6]^{-1}$, where c = 0.69, $R_0$ = 54 Å, and $E_{app} = \frac{I_{Cy5}}{I_{Cy5} + I_{Cy3} \times \frac{(\phi_{Cy5} \times \eta_{Cy5})}{(\phi_{Cy3} \times \eta_{Cy3})}}$. φ and η signify the fluorophores quantum yields and detector channel efficiencies, respectively. The donor and acceptor intensities $I_{Cy3}$ and $I_{Cy5}$, respectively, were corrected for leakage of donor photons into the acceptor channel.

2.2.5 Single-molecule data analysis

Cross-correlation analysis was carried out utilizing customized MATLAB scripts with built-in xcorr function. Time lags for the cross-correlation ranged from 0-5 s (0-50 frames of each 0.1 s integration time). Quality control for the raw smFRET trajectories obtained from the experimental conditions were performed as described[73]. Histograms for data sets measuring pre-mRNA dynamics were constructed by sampling 100 frames of data from each molecule.

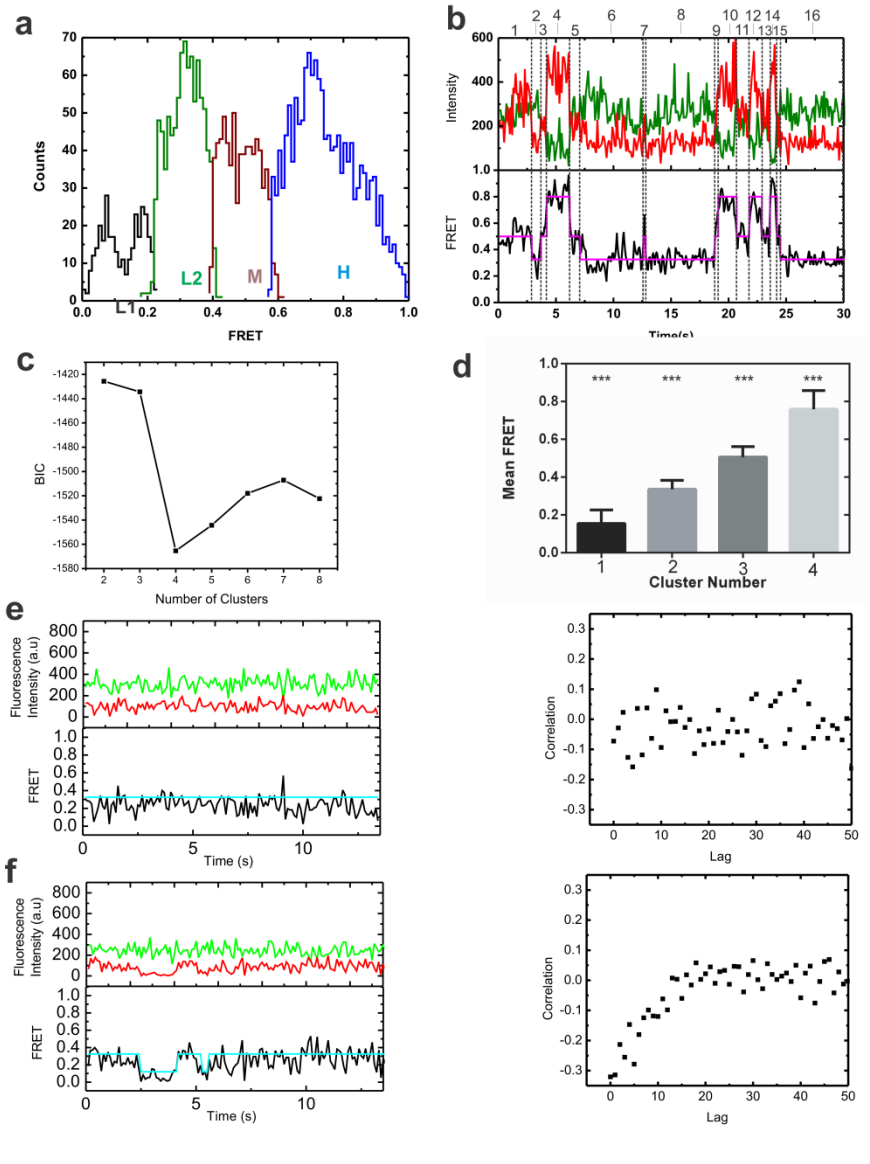



**Figure 2.2 Single molecule clustering and cross correlation analysis**
(a) Histogram of HMM-idealized states for each of the K-means derived clusters L1, L2, M and H obtained by clustering single molecule trajectories from all the experimental conditions ($B^{act}$, B* and C). (b) Representative molecule showing the FRET states assigned by HMM upon K-means clustering. (c) The number of states for clustering was selected through the use of the Bayesian Information Criterion (BIC). (d) The mean and standard deviation of the FRET states from the four K-means derived clusters is shown. *** indicates an extremely significant ($p<0.001$) difference between all pairwise comparisons as determined by the Tukey test. (e) Sample static trajectory from the $B^{act}$ condition with the raw donor (Cy3, green), acceptor (Cy5, red), FRET (black) trajectories, and idealized HMM models (cyan).The corresponding cross correlation analysis of donor and acceptor trajectories with time lags from 0-50 is shown to the right of each panel. (f) Sample dynamic trajectory from the $B^{act}$ condition and corresponding cross correlation analysis.

| Segment | $\mu_{FRET}$ | HMM state | $\mu_{ACP}$-$\mu_{DNR}$/ ($\mu_{ACP}$+$\mu_{DNR}$) | Cluster |
|---|---|---|---|---|
| 1 | 0.51 | 0.5 | 0.039 | M |
| 2 | 0.31 | 0.325 | -0.37 | L2 |
| 3 | 0.51 | 0.5 | 0.030 | M |
| 4 | 0.82 | 0.8 | 0.628 | H |
| 5 | 0.47 | 0.5 | -0.075 | M |
| 6 | 0.32 | 0.325 | -0.352 | L2 |
| 7 | 0.56 | 0.5 | 0.123 | M |
| 8 | 0.32 | 0.325 | -0.353 | L2 |
| 9 | 0.55 | 0.5 | 0.111 | M |
| 10 | 0.76 | 0.8 | 0.525 | H |
| 11 | 0.50 | 0.5 | 0.012 | M |
| 12 | 0.72 | 0.8 | 0.454 | H |
| 13 | 0.52 | 0.5 | 0.061 | M |
| 14 | 0.87 | 0.8 | 0.747 | H |
| 15 | 0.45 | 0.5 | -0.084 | M |
| 16 | 0.31 | 0.325 | -0.366 | L2 |

**Table 2.2 K-means clustering parameters used on the HMM assigned FRET states**
Segment number (first column), the mean raw FRET value (second column), HMM state (third column), and normalized difference of mean acceptor and mean donor intensities for each segment (fourth column).

Histograms for Cwc25-Cy5 and BP-Cy3 FRET experiments were constructed by sampling the entire length of data. Hidden Markov Modeling (HMM) was performed on trajectories utilizing the vbFRET software suite[73]. Each trajectory was individually fit with models ranging from 1-5 states with the optimal number of states determined by the vbFRET algorithm. The inherent experimental variations of the FRET signal between single molecules leads to a slightly different state assignment for similar states across different molecules. A K-means clustering approach was therefore performed in MATLAB to group similar states into larger macro states (L1, L2, M and H). A matrix cataloging the HMM assigned FRET state, raw FRET level, and difference in donor and acceptor intensities for each HMM derived event was utilized as input for the K-means algorithm (**Figure 2.2** and **Table 2.2**). Four macro states were identified whose boundaries were used to re-assign the original HMM idealized FRET states. The number of macro states was determined using the Bayesian Information Criterion (BIC) as a model selection tool. K-means and BIC have been used previously to group and accurately determine the number of clusters in single molecule data[78-80]. The BIC score was calculated in MATLAB as follows:

$$BIC = (-2 * LLF) + (NumParams * log(NumObs))$$

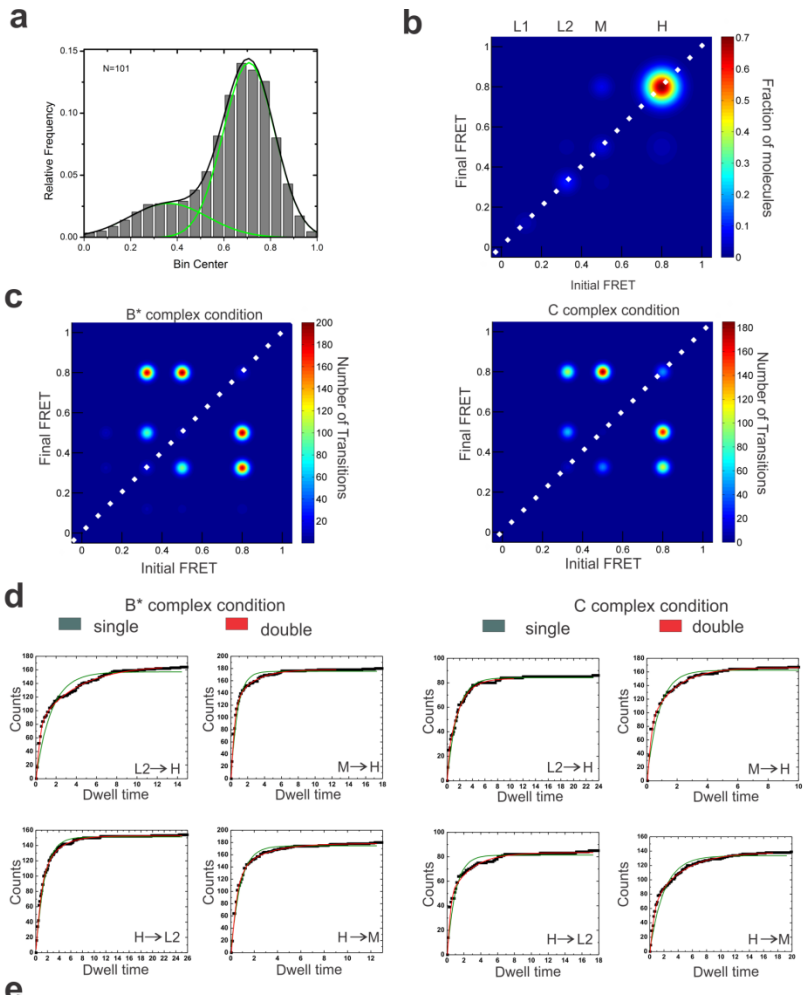
where LLF is the log-likelihood function, NumParams the number of parameters, and NumObs the number of observations. The number of clusters (K) was varied from 2-8. The lowest BIC score was achieved for k = 4 (**Figure 2.2c**). Once identified, the clusters were tested for similarity using the multi-comparison Tukey test in the software package PRISM-GraphPad. The results of this analysis indicate that there is a significant difference ($p<0.001$) for the pairwise comparison of all four clusters (**Figure 2.2d**).

Transition Occupancy Density Plots (TODPs) were used to plot the fraction of molecules that contain any given HMM transition at least once[73]. Molecules that did not exhibit any transitions were plotted along the diagonal at their respective positions. For kinetic rate calculations, Transition Density Plots (TDPs) that are scaled by the number of times a transition occurs irrespective of how many molecules exhibit that transition were used as described[73]. A cumulative histogram scatter plot was then fit with a double-exponential association equation in MicroCal Origin (**Figure 2.3**). A weighted average ($k_{w,observed}$) of the two rate constants from the double-exponential fits was calculated based on the amplitude values of the exponential equation. To correct for bias introduced by the limited observation window used to measure dwell times, the measured $k_{w,observed}$ values were corrected by subtracting the photobleaching rate constant and the reciprocal of the observation window to yield $k_{w,actual}$ as described[81]. Equilibrium constants ($K_{eq}$) were calculated by taking a ratio of the forward and backward rate constants for a set of state-to-state transitions. Post-synchronized histograms (PSHs) were constructed by synchronizing individual FRET events to the time where one of the macro-states (M, H) was achieved. The scale bar represents the fraction of FRET events which exhibit a certain FRET state at a given time.

## 2.3 Results

### 2.3.1 Purifying $B^{act}$ in complex with FRET-labeled Ubc4 pre-mRNA

It has been known for almost 25 years that a Prp2-1 yeast splicing extract can be heat inactivated[37]. In this extract, the spliceosome is fully assembled but cannot carry out the first step of splicing. The immature Prp2-1 spliceosome purified by gradient centrifugation ($B^{act}$) was shown to proceed through the first step of splicing only upon the addition of Prp2 protein and heat-stable factor(s)[68]. More recently, this experiment was repeated with




e

| | L2 -> H | | H -> L2 | | M -> H | | H -> M | |
|---|---|---|---|---|---|---|---|---|
| Parameter | B* | C | B* | C | B* | C | B* | C |
| A1 | 90.50 +/- 1.32 | 70.53 +/- 1.10 | 65.60 +/- 4.30 | 40.58 +/- 1.23 | 61.00 +/- 3.03 | 91.65 +/- 2.08 | 54.90 +/- 4.14 | 80.33 +/- 1.18 |
| $t_1$ (s) | 3.92 +/- 0.14 | 1.84 +/- 0.05 | 2.82 +/- 0.11 | 2.54 +/- 0.10 | 2.33 +/- 0.13 | 1.65 +/- 0.05 | 2.40 +/- 0.18 | 4.09 +/- 0.09 |
| A2 | 94.40 +/- 2.75 | 27.76 +/- 5.61 | 89.70 +/- 4.01 | 55.51 +/- 4.01 | 135.84 +/- 3.24 | 105.75 +/- 3.40 | 140.41 +/- 4.10 | 68.46 +/- 1.59 |
| $t_2$ (s) | 0.29 +/- 0.02 | 0.08 +/- 0.03 | 0.78 +/- 0.04 | 0.21 +/- 0.02 | 0.35 +/- 0.02 | 0.16 +/- 0.01 | 0.43 +/- 0.02 | 0.45 +/- 0.02 |
| $t_w$ (s) | 2.07 | 1.34 | 1.64 | 1.19 | 0.96 | 0.85 | 0.98 | 2.42 |
| $k_{w,observed}$ ($s^{-1}$) | 0.48 | 0.74 | 0.61 | 0.84 | 1.04 | 1.17 | 1.01 | 0.41 |
| $k_{w,actual}$ ($s^{-1}$) | 0.42 | 0.68 | 0.55 | 0.78 | 0.98 | 1.11 | 0.95 | 0.35 |

**Figure 2.3 Post-first step splicing signature and single molecule kinetic analysis**
(a) FRET probability distribution for molecules stalled by the addition of Prp16 (K379A) dominant negative (DN) mutant. (b)TODP for Prp16DN mutant. (c) Transition Density Plots (TDPs) for the B* and C complex molecules scaled to the number of transitions determined by HMM. (d) Cumulative distribution plot of dwell times extracted for the indicated transition and fit with either a single- or double-exponential rate equation. (e) Parameters for the double-exponential equations fitted to the dwell time data. To reduce the dimensionality of the data, a weighted average rate constant $k_w$ was calculated by utilizing the amplitudes associated with each time constant as weighting factors. $k_w$ was used for $K_{eq}$ calculations and rate comparisons between B* and C complex conditions.

purified Prp2, Spp2 and the (since identified) heat stable factor Cwc25[38]. Such a purified system is ideal for exploring substrate dynamics using single-molecule FRET. To this end, we constructed a yeast strain containing the *prp2-1* mutation and a tandem affinity purification (TAP) tag derivative of one of the NTC components, Cef1, known to be present in the spliceosome at this stage[38,39]. This approach allowed us to purify the stalled $B^{act}$ complex via the Cef1-TAP tag using biotin-IgG bound to streptavidin.

The Ubc4 pre-mRNA used in this study was synthesized chemically and labeled with the FRET donor Cy3, which is 6 nucleotides downstream from the BP adenosine, and with the FRET acceptor Cy5, which is 7 nucleotides upstream from the 5'SS (**Table 2.1**). Splicing reactions were assembled for 30 minutes using heat-inactivated Prp2-1 extract containing the fluorophore-labeled pre-mRNA in the presence of 2 mM ATP. We bound the pre-catalytic $B^{act}$ spliceosome either to streptavidin-coated magnetic beads (for biochemical analysis) or to a PEG-passivated slide coated with streptavidin, biotin-IgG, and excess free biotin (for single molecule analysis) (**Figure 2.4a**). Free biotin was added to block any biotin binding sites not associated with biotin-IgG and prevent direct binding of the 5' biotinylated pre-mRNA to the slide. Surface binding of the pre-mRNA alone was at least 11-fold lower than that of the Cef1-TAP-tagged $B^{act}$ complex containing the pre-mRNA (**Figure 2.1a,b,d**). We detected similarly minimal nonspecific binding of the Cef1-tagged complex when we omitted biotin-IgG **(Figure 2.1c)** or when a TAP tag was present on Prp4, a protein that was recently shown to be absent from the $B^{act}$ complex (data not shown)[39]. To verify that our purification yielded functional $B^{act}$ pre-spliceosome, we added micrococcal nuclease (MNase)-treated (and thus RNA-free) whole cell extract to the stalled, bead-purified $B^{act}$ complex and incubated the mixture for 40 min at 23 ºC under splicing conditions. We found that both steps of splicing were reconstituted (**Figure 2.5**),

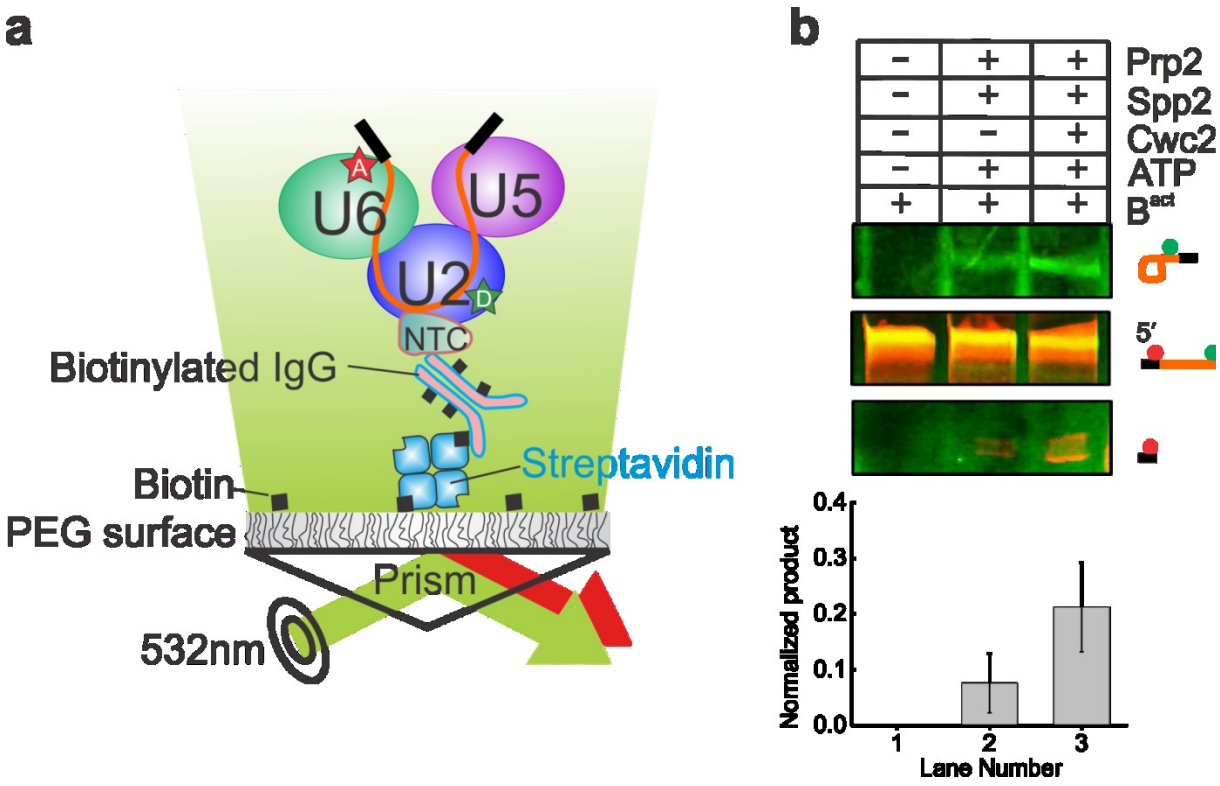



**Figure 2.4 The SiMPull-FRET approach used to interrogate active splicing complexes**
(a) Schematic showing the affinity purified $B^{act}$ complex immobilized to a streptavidin coated quartz slide via biotinylated-IgG. The green and red stars on the pre-mRNA represent Cy3 and Cy5 fluorophores, respectively. (b) A 15% urea-polyacrylamide gel scanned using a variable mode Typhoon imager shows pre-mRNA and first-step products (the top panel shows intron-lariat, the middle panel shows pre-mRNA, and the bottom panel shows the 5'-exon, in each case rendered using an overlay of the Cy3 and Cy5 scans). Error bars indicate the standard deviation obtained from triplicate experimental sets.

demonstrating that only proteins are needed to chase $B^{act}$ into splicing the substrate. In addition, we bead-purified the $B^{act}$ complex and found that the recombinant proteins Prp2 and Spp2 (**Figure 2.6**) are sufficient to yield an amount of first-step splicing products (**Figure 2.4b**) that is slightly higher than that observed for the previously characterized actin pre-mRNA[38] (**Figure 2.5b,c**). Additionally when we added Cwc25, we observed a two- to ten-fold enhancement of first-step splicing (**Figure 2.4b**). Our fluorophore-labeled Ubc4 construct showed a first-step splicing efficiency ranging from 12% to 40% when incubated at 23 °C for 30-40 min, within two-fold of that of other labeled and unlabeled Ubc4 constructs[53]. Taken together, these experiments establish that FRET labeling the pre-mRNA substrate is compatible with the expected assembly and splicing activity of the immunopurified $B^{act}$ complex, paving the way for SiMPull-FRET interrogation.

2.3.2 $B^{act}$ complex holds the pre-mRNA 5'SS and BP in a distal conformation

To obtain mechanistic insight into pre-mRNA splice-site juxtaposition during the Prp2-driven restructuring of the $B^{act}$ complex into the catalytically activated B* complex, we carried out SiMPull-FRET on the slide-bound $B^{act}$ complex (**Figure 2.4a**) in standard splicing buffer. After verifying that each selected pre-mRNA molecule contained one Cy3 and one Cy5 fluorophore, we collected FRET values over the first 100 video frames (at 100-ms time resolution) from 297 molecules. Histograms of the FRET values indicated a single Gaussian distribution with an average FRET value of 0.3±0.15 (s.d.) (**Figure 2.7a**). Given the background noise inherent to any single-molecule experiment, we used hidden Markov modeling (HMM) to find the underlying FRET states[73]. Using a K-means approach, we then clustered the HMM-assigned states using all the experimental conditions from this study into four macro states with FRET values of 0.0-0.23 (state L1), 0.23-0.42 (L2), 0.42-0.60 (M), and 0.6-1.0 (H) (**Figure 2.2** and

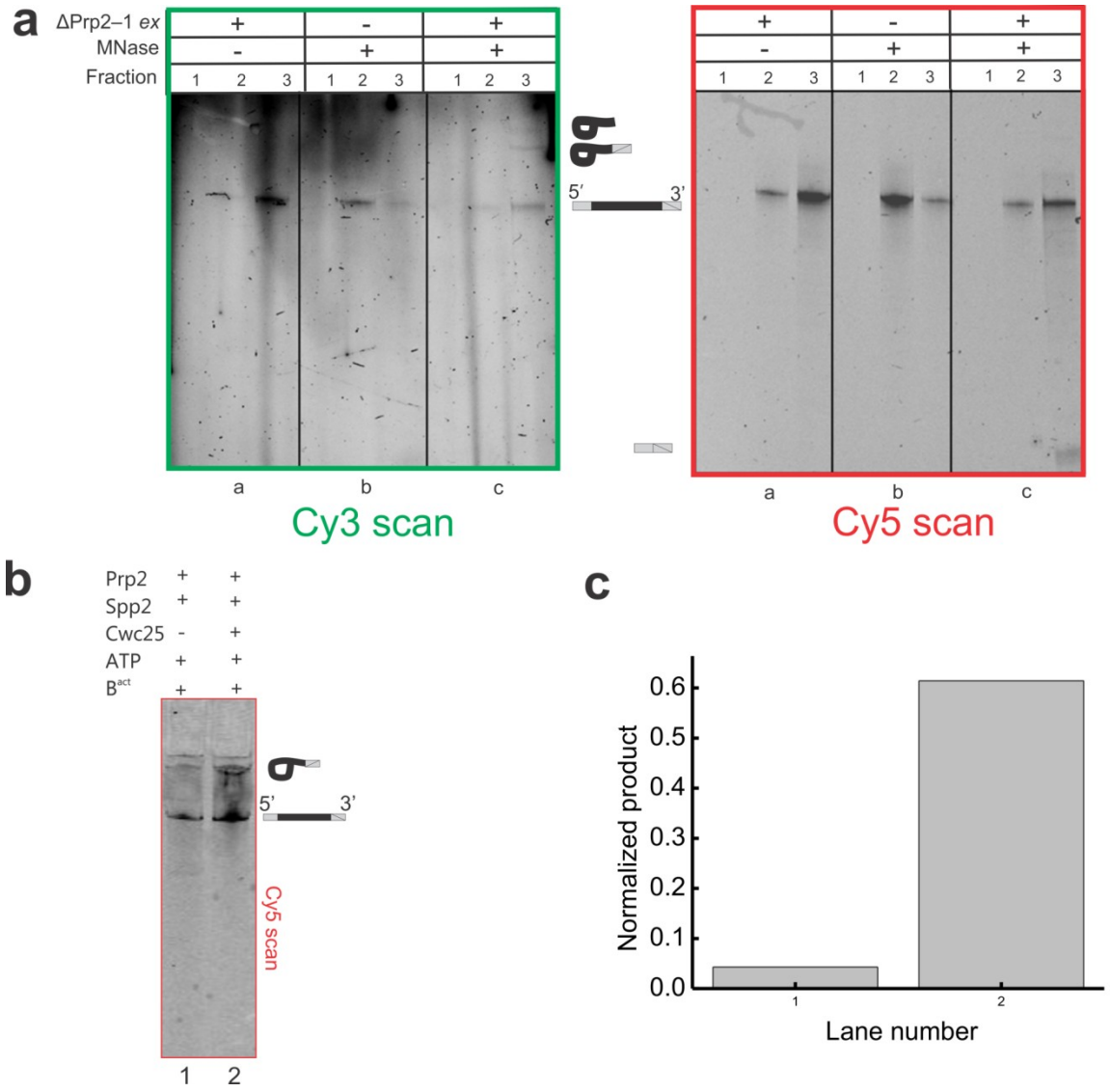



**Figure 2.5 Confirmation of $B^{act}$ complex specificity and activity**
(a) 15% Urea-polyacrylamide gel scanned with a variable mode Typhoon imager. The intron and intron-lariat products are observed in the Cy3 scan (left) and the mature mRNA product is visualized in the Cy5 scan (right). Lanes 1, 2 and 3 represent fractions wash, unbound and bound, respectively. Conditions a and c represent wild-type Ubc4 pre-mRNA assembled in $B^{act}$ complex and immobilized on magnetic beads with biotin-IgG. Condition b is wild-type pre-mRNA assembled in the absence of extract. Bound molecules were reconstituted with or without Micrococcal Nuclease (MNase) treated extract. (b) 6% Urea-polyacrylamide gel scanned using a variable mode Typhoon imager. Affinity purified $B^{act}$ complex formed with Cy5-actin pre-mRNA supplemented with Prp2, Spp2 and 2 mM ATP (lane 1) and Prp2, Spp2, Cwc25 (lane 2). (c) Quantification of lanes 1 and 2 from panel b.

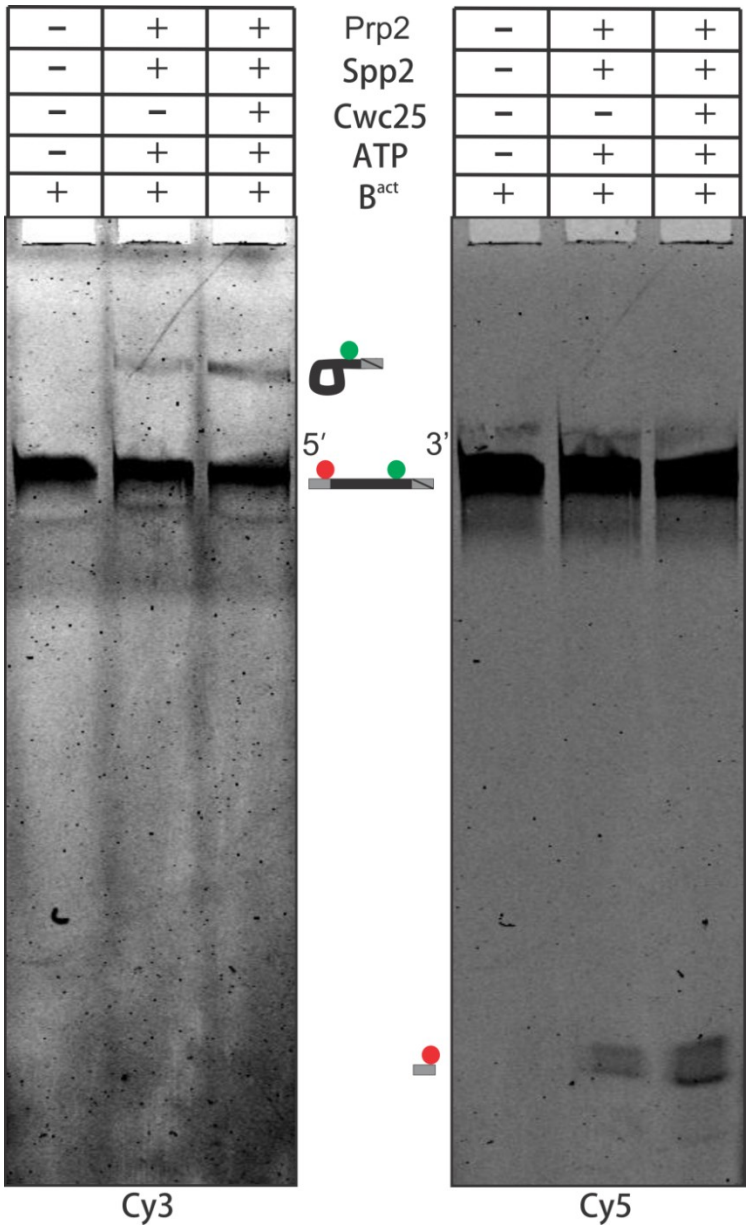



**Figure 2.6 Confirmation of $B^{act}$ complex activity using recombinant proteins**
15% polyacrylamide gel scanned with a variable mode Typhoon imager. Ubc4 pre-mRNA assembled in $B^{act}$ complexes and supplemented with or without recombinant proteins Prp2, Spp2, Cwc25. This represents the uncropped unedited form of the gel presented in Figure 2.4

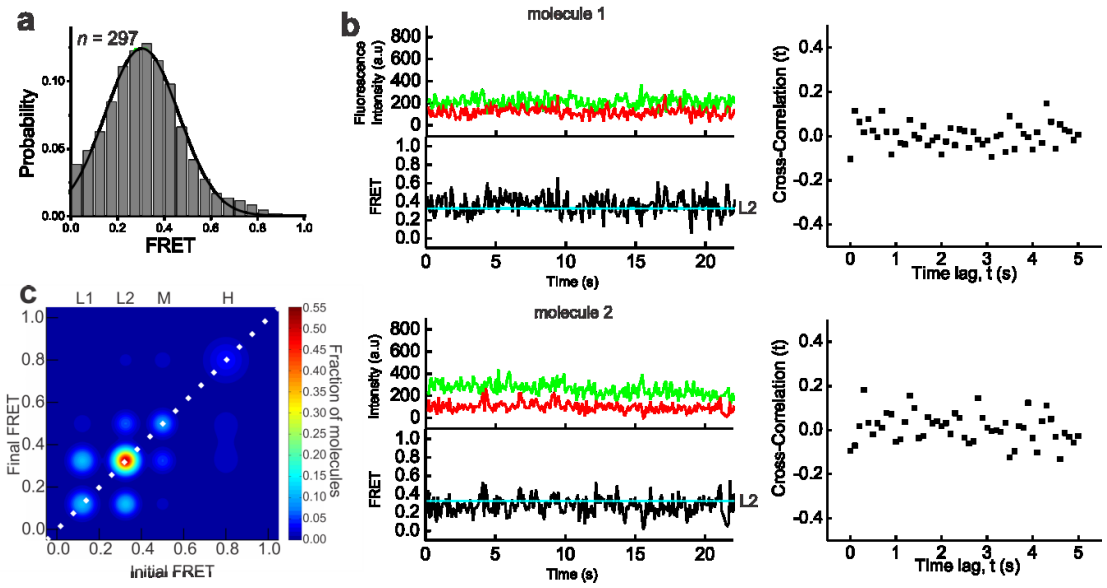



**Figure 2.7 In the Prp2-stalled $B^{act}$ complex, the pre-mRNA is predominantly restricted to a static low FRET state**

(a) FRET probability distribution of the raw single molecule FRET trajectories from the purified $B^{act}$ complex. (b) Representative time traces of the $B^{act}$ complex with raw donor (Cy3, green), acceptor (Cy5, red), and FRET (black) trajectories and their idealized HMM (cyan). The right panel of each molecule is the corresponding cross-correlation of donor and acceptor intensities. (c) Transition Occupancy Density Plot (TODP) showing the fraction of $B^{act}$ complex molecules that either lack a transition and thus lie on the diagonal (dotted white line) or transition from one indicated FRET state to another. L1, L2, M, or H refers to the four states resulting from clustering analysis.

**Table 2.2**). The dominant behavior in the $B^{act}$ complex is a static low-FRET state L2 (**Figure 2.7b**). Transition Occupancy Density Plots (TODPs), which are scaled to emphasize the transitions found to be most common among a molecule population[73], indicate that the static L2 state represents the only behavior in ~52% of all $B^{act}$ molecules (**Figure 2.7c** and **Table 2.3**). In addition, molecules in this state have few transitions (**Figure 2.7b**, HMM fit, cyan line). To test for dynamics that would be too fast for detection by HMM, we performed cross-correlation analysis between the donor and acceptor trajectories of each molecule, and in the resulting scatter around 0, found no evidence for rapid transitions (**Figure 2.7b** and **Figure 2.2**). Although splice site recognition begins in the splicing cycle as early as the commitment complex[82], our results suggest that the 5'SS and BP in the $B^{act}$ complex are kept stably apart, probably not close enough for splicing chemistry to occur.

2.3.3 Prp2 mediates an NTP-dependent remodeling of the pre-mRNA

The ATPase action of Prp2 has been shown to catalyze a large conformational change that activates the spliceosome for the first step of splicing[38,39,68]. Spliceosomal binding of Prp2 is dependent on its interaction with the G patch domain of its cofactor protein Spp2[83,84]. The addition of Prp2, Spp2 and ATP transforms the pre-catalytic $B^{act}$ complex into the catalytically active, distinctly sedimenting B* complex and results in low levels of first-step splicing[38]. To investigate the role of Prp2 in pre-mRNA remodeling during this step, we incubated the $B^{act}$ complex assembled on the slide surface with Prp2, Spp2 and 2 mM ATP (henceforth referred to as B* complex conditions). B* complex conditions resulted in a substantial shift in the FRET histogram toward a new ~45% population with a mean FRET value of 0.71±0.01 (s.d.), diminishing the lone 0.33±0.01 FRET distribution observed for the $B^{act}$ complex (**Figure 2.8a**). In contrast to the predominantly static L2 state of the $B^{act}$ complex, molecules under B*

| Initial FRET | Final FRET | Fraction of molecules | | | Prp16 DN |
|---|---|---|---|---|---|
| | | Bact | B* | C | |
| 0.12 | 0.12 | 0.11 | 0.00 | 0.00 | 0.03 |
| 0.12 | 0.33 | 0.11 | 0.02 | 0.00 | 0.00 |
| 0.12 | 0.50 | 0.02 | 0.01 | 0.00 | 0.00 |
| 0.12 | 0.80 | 0.01 | 0.02 | 0.00 | 0.00 |
| 0.33 | 0.12 | 0.12 | 0.02 | 0.01 | 0.00 |
| 0.33 | 0.33 | 0.52 | 0.11 | 0.06 | 0.09 |
| 0.33 | 0.50 | 0.06 | 0.12 | 0.08 | 0.02 |
| 0.33 | 0.80 | 0.01 | 0.39 | 0.29 | 0.01 |
| 0.50 | 0.12 | 0.01 | 0.01 | 0.00 | 0.00 |
| 0.50 | 0.33 | 0.05 | 0.11 | 0.08 | 0.02 |
| 0.50 | 0.50 | 0.09 | 0.06 | 0.03 | 0.06 |
| 0.50 | 0.80 | 0.02 | 0.26 | 0.31 | 0.07 |
| 0.80 | 0.12 | 0.01 | 0.02 | 0.00 | 0.00 |
| 0.80 | 0.33 | 0.01 | 0.37 | 0.30 | 0.01 |
| 0.80 | 0.50 | 0.01 | 0.24 | 0.29 | 0.03 |
| 0.80 | 0.80 | 0.04 | 0.12 | 0.32 | 0.70 |

**Table 2.3 TODP quantification for all data sets**
Molecules with at least one occurrence of the FRET transition given by the Initial and Final FRET states in columns one and two are counted and divided by the total number of molecules in that transition. Molecules that only occupy one state are accounted for in rows where the Initial and Final FRET states are equal.

conditions show dynamic (reversible) excursions to high-FRET states, indicating that the 5'SS and BP can now reach the close proximity required for first-step chemistry. More specifically, the B* condition comprises the L2, M and H FRET states, where the H state is accessed from either the L2 or M states (**Figure 2.8b**). TODP plots show that only ~11% of molecules retain the static L2 state characteristic of the $B^{act}$ complex, whereas ~39% of molecules exhibit at least one L2-to-H transition (**Figure 2.8c** and **Table 2.3**). Notably, transitions into the H state were short lived in the majority of molecules (**Figure 2.8b**). However, 12% of molecules showed a static high-FRET state (**Figure 2.8c**), indicating that they made a transition through the low levels of first-step splicing observed under B* conditions, after which the labeled 5'SS and BP become covalently linked (**Figure 2.4b**). To verify that this static high-FRET state corresponds to the pre-mRNA substrate configuration after the first step of splicing, we trapped this configuration using a dominant-negative Prp16 mutant (Prp16DN; K379A)[39,85,86] added to the (non-heat-inactivated) Cef1-TAP-tagged Prp2-1 yeast extract in the presence of 2 mM ATP. This protocol is expected to enrich for the post-first-step C complex, which was then immobilized on the slide surface, washed and imaged. The resulting histogram showed a dramatic enrichment to ~76% of a high-FRET population with a mean FRET value of 0.7±0.01 (s.d.) (**Figure 2.3** and **Table 2.3**). TODP analysis revealed that ~70% of all molecules adopt the same static high-FRET state first observed under B* conditions, strongly supporting the notion that these molecules indeed have undergone the first chemical step of splicing.

Prp2 can directly bind a region in the pre-mRNA downstream of the BP adenosine, even in the absence of ATP[87,88]. To investigate whether Prp2 alone can induce the observed pre-mRNA remodeling, we omitted Prp2 or Spp2 from our B* conditions and found the resulting FRET histograms to be indistinguishable from those of the starting $B^{act}$ complex (**Figure 2.8d**,

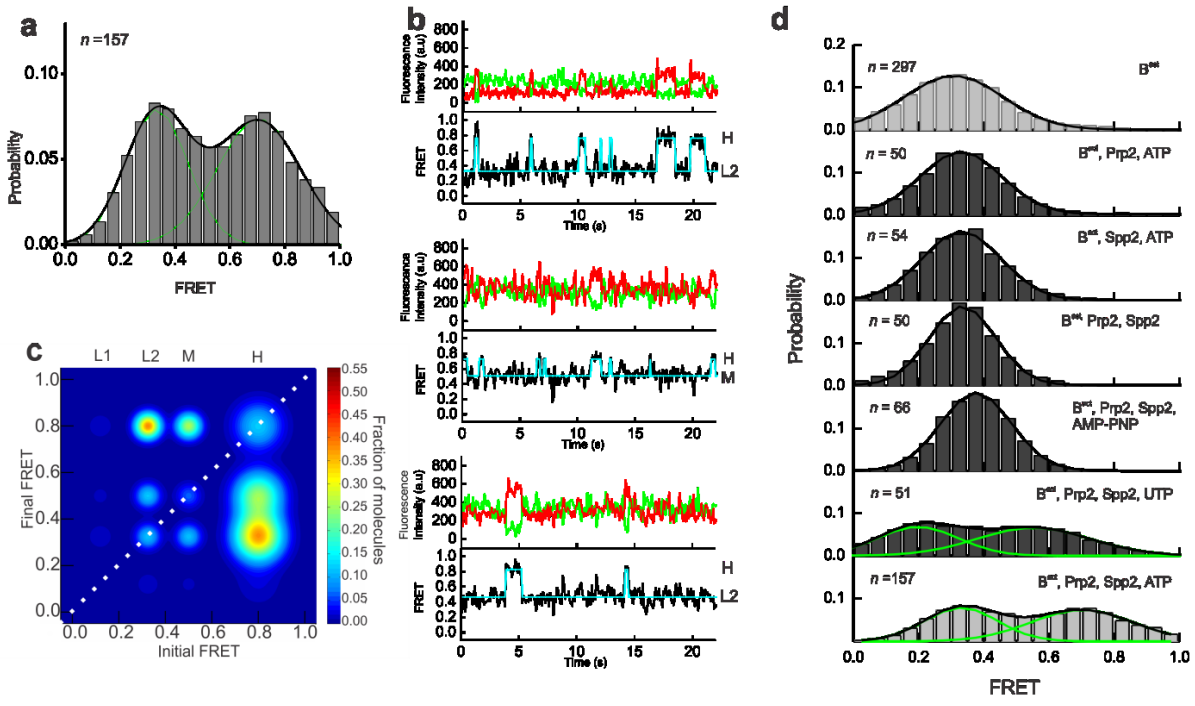



**Figure 2.8 Upon the addition of ATP, Prp2, and Spp2, the pre-mRNA is able to explore splice site proximity**

(a) FRET probability distribution of the raw single molecule FRET trajectories upon the addition of Prp2, Spp2, and ATP to the $B^{act}$ complex, leading to formation of the B* complex. (b) Representative single molecule FRET time trajectories from the B* condition with the raw donor (Cy3, green), acceptor (Cy5, red), FRET (black) trajectories, and idealized HMM models (cyan). (c) TODP generated from the idealized HMM for molecules in the B* condition. L1, L2, M, or H refers to the four states resulting from clustering analysis. (d) FRET probability densities generated from molecules in $B^{act}$ incubated with various combinations of components required for formation of B* (Prp2, Spp2, and ATP). In addition, $B^{act}$ was incubated with Prp2, Spp2, and one of two NTP analogs, non-hydrolysable AMP-PNP or UTP.

average FRET value of 0.31±0.16 and 0.33±0.13 (s.d.), respectively). Next, we studied the role of ATP in the remodeling of the $B^{act}$ complex. For most spliceosomal DExD/H-box helicases, both ATP-dependent and ATP-independent roles have been proposed[89-93]. Prp2 in particular has been shown to cause extensive conformational remodeling of the spliceosome in the absence of ATP, whereas the displacement of SF3b is ATP dependent[39]. To determine whether the pre-mRNA remodeling observed here requires ATP, we incubated the $B^{act}$ complex with Prp2 and Spp2 in the absence of ATP and observed no appreciable change in the FRET histogram (**Figure 2.8d**, average FRET value of 0.33±0.12 (s.d.)). When the non-hydrolysable ATP analog AMPPNP was used instead of ATP, the FRET histogram was again similar (average FRET value of 0.37±0.11) to that of the $B^{act}$ complex, with no notable excursion to higher FRET states, showing that ATP hydrolysis is required for these excursions to occur (**Figure 2.8d**). (We note that the minor upwards shift observed in the histogram may be due to binding of AMPPNP to Prp2, resulting in a slight conformational change.) Finally, the DExD/H box helicases involved in spliceosomal reorganization are either integral components of snRNPs or extrinsic components, as is Prp2. MS studies have shown that the stalled $B^{act}$ complex contains stoichiometric amounts of the DExD/H-box helicase Brr2, the integral component of the U5 snRNP responsible for U4-U6 unwinding[94]. Direct interactions between Prp2 and the C-terminus of Brr2 have recently been discovered[88], suggesting a possible role for Brr2 in first-step catalytic activation. To test this possibility, we exploited the fact that Brr2 is a strict ATPase[95] whereas Prp2 is a broad NTPase[96], and supplemented the $B^{act}$ complex under B* conditions with UTP instead of ATP. The resulting FRET efficiency histogram is clearly distinct from that of $B^{act}$ and overlays well with that of the ATP-mediated B* condition, with a slightly less-efficient shift toward the higher FRET population (**Figure 2.8d**). This lower efficiency is consistent with the

~2-fold reduction in activity of Prp2 in the presence of NTPs other than ATP[96]. Collectively, these results indicate that the NTP-driven helicase activity of Prp2 in complex with its activator Spp2 causes a large structural reorganization of the pre-mRNA that allows the distal 5'SS and BP of the $B^{act}$ complex to reversibly access proximal conformations, which in turn enable first-step splicing.

2.3.4 Cwc25 enhances first-step splicing by H state stabilization

Although the pre-mRNA is remodeled by the NTPase action of Prp2–Spp2, it does not undergo efficient first-step catalysis. Further enhancement of first-step splicing efficiency requires addition of Cwc25 to the $B^{act}$ complex incubated with Prp2, Spp2 and ATP (**Figure 2.4b**). Cwc25 was identified as one of a group of proteins in complex with Cef1–Ntc85 of the NTC complex[97]. To determine the role of Cwc25 in remodeling of the pre-mRNA, we performed SiMPull-FRET on the purified $B^{act}$ complex supplemented with Prp2, Spp2, ATP and Cwc25 (henceforth referred to as C complex conditions). This resulted in a FRET histogram with an enhanced ~73% population with a mean FRET value of 0.75±0.01 (s.d.) (**Figure 2.9a**). We found the FRET states under C complex conditions to be the same as those under B* conditions, with the L2-to-H and M-to-H transitions prevalent; however, the occupancy in the H state was considerably enhanced under C conditions (**Figure 2.9b**). TODP analysis revealed the fraction of molecules displaying at least one L2-to-H and M-to-H transition to be similar under B* and C conditions, whereas the static H state occupancy (35%) was ~3-fold increased (**Figure 2.9c** and **Table 2.3**). This shift is similar in magnitude to the enhancement in first-step splicing induced by Cwc25 (**Figure 2.4b**), consistent with the static H state representing the pre-mRNA in the C complex after the first chemical step of splicing. Post-synchronized histograms (PSHs) created by aligning the HMM-fitted traces to start at the M state show that the molecules under C

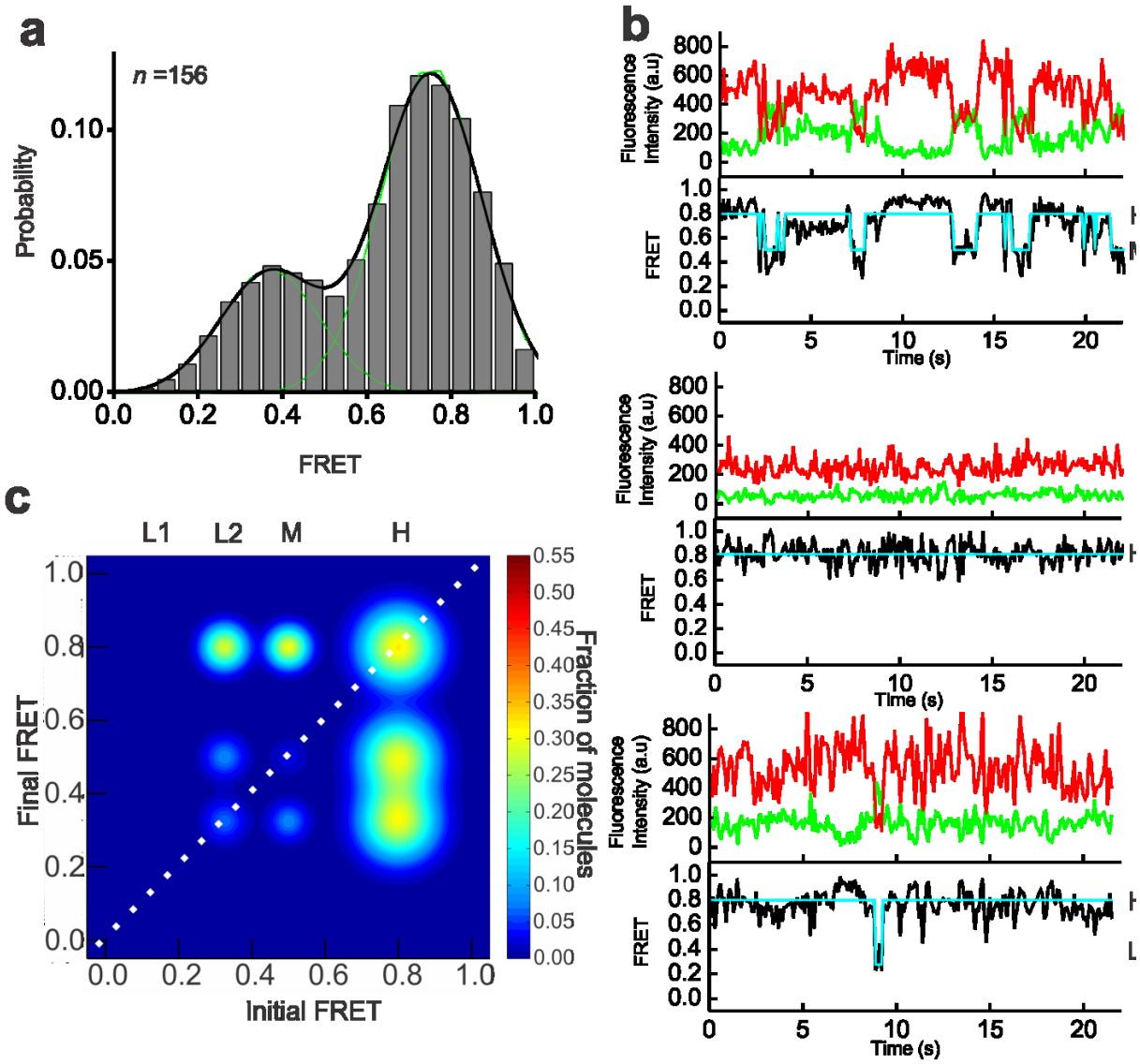



**Figure 2.9 Under C complex conditions, the pre-mRNA accesses dynamic and stabilized high-FRET states**

(a) FRET probability distribution of the raw single molecule FRET trajectories upon the addition of Prp2, Spp2, Cwc25, and ATP to the $B^{act}$ complex. (b) Representative single molecule FRET trajectories of molecules from the C complex condition with the raw donor (Cy3, green), acceptor (Cy5, red), FRET (black) trajectories, and idealized HMM models (cyan). (c) TODP generated from the idealized HMM for molecules in the C complex condition. L1, L2, M, or H refers to the four states resulting from clustering analysis.

conditions both transition more frequently to the H state and exhibit a higher residence time once in the H state (**Figure 2.10a**). A similar comparison of transitions starting at the H state further emphasizes the stabilization of this state by Cwc25 under C conditions (**Figure 2.10a**). To rule out that a change in photostabilty of molecules in the C complex affects the relative prevalence of the H state, we analyzed the average photobleaching time under $B^{act}$, B* and C conditions and found them to be comparable (**Table 2.4**). To quantitatively characterize the effects of Cwc25 on the conversion of the B* to the C complex, we plotted the cumulative dwell times for the forward and backward L2-to-H and M-to-H transitions under both conditions and fit them with double-exponential functions (**Figure 2.3**). A comparison of the weighted average rate constant for the L2-to-H transition showed similar forward and backward rate constants under both conditions, yielding equivalent equilibrium constants $K_{eq} = k_{forward}/k_{backward}$ of ~0.80 (**Figure 2.3c**). In contrast, the presence of Cwc25 accelerates the forward and reduces the backward rate constant of the observed M-to-H transition, leading to a $K_{eq}$ that is ~3-fold more favorable for the dynamic H state under C conditions than under B* conditions (**Figure 2.10b)**. Notably, the molecules in the static H state, which results from the chemical bond formed after first-step catalysis, do not contribute to this kinetic effect. State dwell times cannot be calculated for such molecules, which are only in one state of poorly defined duration during the entirety of our observation window. The effect of static H molecules is therefore more appropriately represented by the increase in molecules of high FRET on the TODP diagonal (compare **Figure 2.8** and **Figure 2.9**). We also note that both dynamic and static H state molecules, however, do contribute to the enhanced high-FRET peak of the histogram in **Figure 2.9a.**

To show that the same pre-mRNA molecule can be converted from B* to C complex, we observed the same field of view before and after shifting from B* conditions (excluding Cwc25)

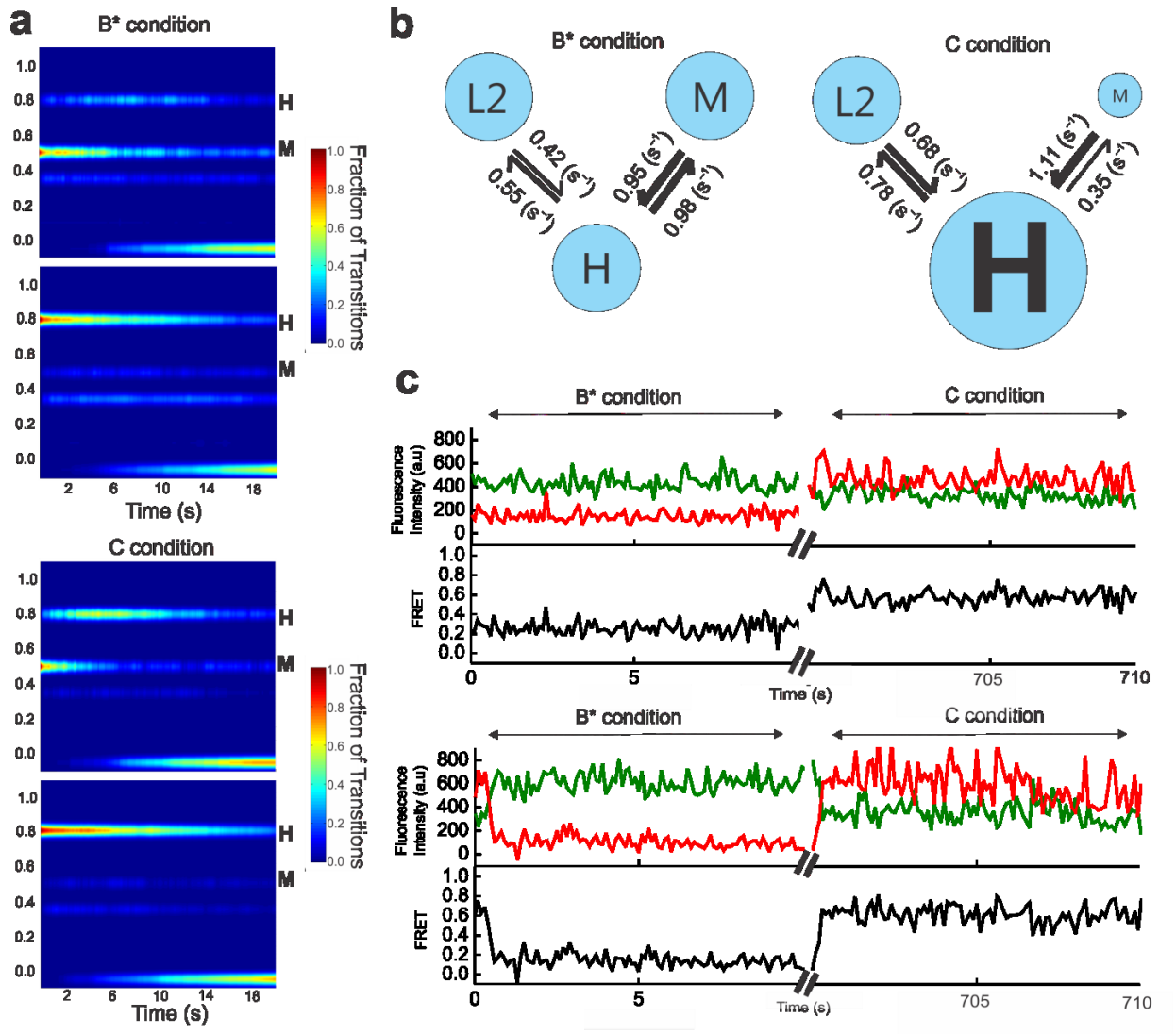



**Figure 2.10 Cwc25 enhances the first step of splicing by stabilizing the H state**
(a) A comparison of the aggregate molecular behavior before (B* condition) and after (C condition) Cwc25 addition through post-synchronized histograms (PSH) with all trajectories synchronized to start from either the M (top) or H state (bottom). (b) A comparison of the rate constants of the observable transitions under B* and C conditions. The thickness of arrows corresponds to the relative rate constants. (c) Representative single molecule FRET trajectories showing transition dynamics from the same molecules imaged before (B* condition) and after (C condition) Cwc25 addition with the raw donor (Cy3, green), acceptor (Cy5, red), and FRET (black) trajectories. The axis breaks represent 10 min of incubation after Cwc25 addition.

| Condition | Number of molecules | Average photobleaching time (Seconds) |
|---|---|---|
| $B^{act}$ complex | 297 | 11.8±7.0 |
| B* complex | 157 | 15.9±10.9 |
| C complex | 154 | 16.5±13.5 |

**Table 2.4 Comparison of average photobleaching times and number of molecules per condition**

to C conditions (including Cwc25) and incubating for 10 min in the dark. Before the dark period, molecules were dynamically shuttling between the L2 and H states. A subset of molecules were observable after the dark period and of those, ~50% shifted to the stabilized H state (**Figure 2.10c** and **Table 2.5**). Taken together, our results suggest that Cwc25 acts kinetically to stabilize the catalytically favorable conformation, thereby effecting an enhancement of the first chemical step of splicing.

2.3.5 Cwc25 dynamically interacts near the BP upon B* formation

Previous studies have shown that Cwc25 binds stably to the spliceosome after Prp2-mediated SF3a–SF3b destabilization[40]. It seems likely that Cwc25 enhances first-step chemistry by binding to the pre-mRNA, as mutation at the BP abolishes this interaction[41]. Cwc25 was also recently shown to cross-link near the BP of the pre-mRNA[98]. To directly observe the binding of Cwc25 to the pre-mRNA, we labeled the protein's C-terminus with Cy5. The Cy5 near the 5'SS of the pre-mRNA was pre-bleached so that the pre-mRNA had a single fluorescent Cy3 label near the BP. We tested the activity of the Cy5-tagged Cwc25 (Cwc25-Cy5) using our bead pulldown assay and found it to be fully functional. SiMPull-FRET experiments were then carried out with Cwc25-Cy5 added to the $B^{act}$ complex with and without Prp2, Spp2 and ATP (**Figure 2.11a**). We observed repeated binding and dissociation of Cwc25-Cy5 and resulting FRET with the BP under B* condition (**Figure 2.11b,c**). On the basis of the FRET distribution of these binding events, centered around 0.37±0.03 (s.d.), we estimate that Cwc25 binds within FRET distance of the Cy3-Cy5 pair (<100 Å), roughly ~52 Å from the BP adenosine (**Figure 2.11b**). The lower FRET peak, centered around 0.15, represents background signal. By contrast, there was little observable FRET between Cwc25 and the pre-mRNA BP in the absence of Prp2–Spp2

| Fraction | Pre-Cwc25 addition | Post-Cwc25 addition | Classification |
|---|---|---|---|
| 0.11 | High FRET with long dwell | High FRET with long dwell | Catalysis of 1st step |
| 0.39 | Various behaviors (excluding high FRET with long dwell) | High FRET with long dwell | Cwc25 mediated enhancement |
| 0.33 | Fast Switching | Slow/No witch (no stable high FRET) | Prp2 unbound |
| 0.08 | Fast Switching | Fast Switching | Cwc25 not bound |
| 0.09 | N/A | Elevated Noise | N/A |

**Table 2.5 Classification of molecules from the observation of the same molecule chased from the $B^{act}$ to the C complex with the inclusion of a dark period during Cwc25 addition**

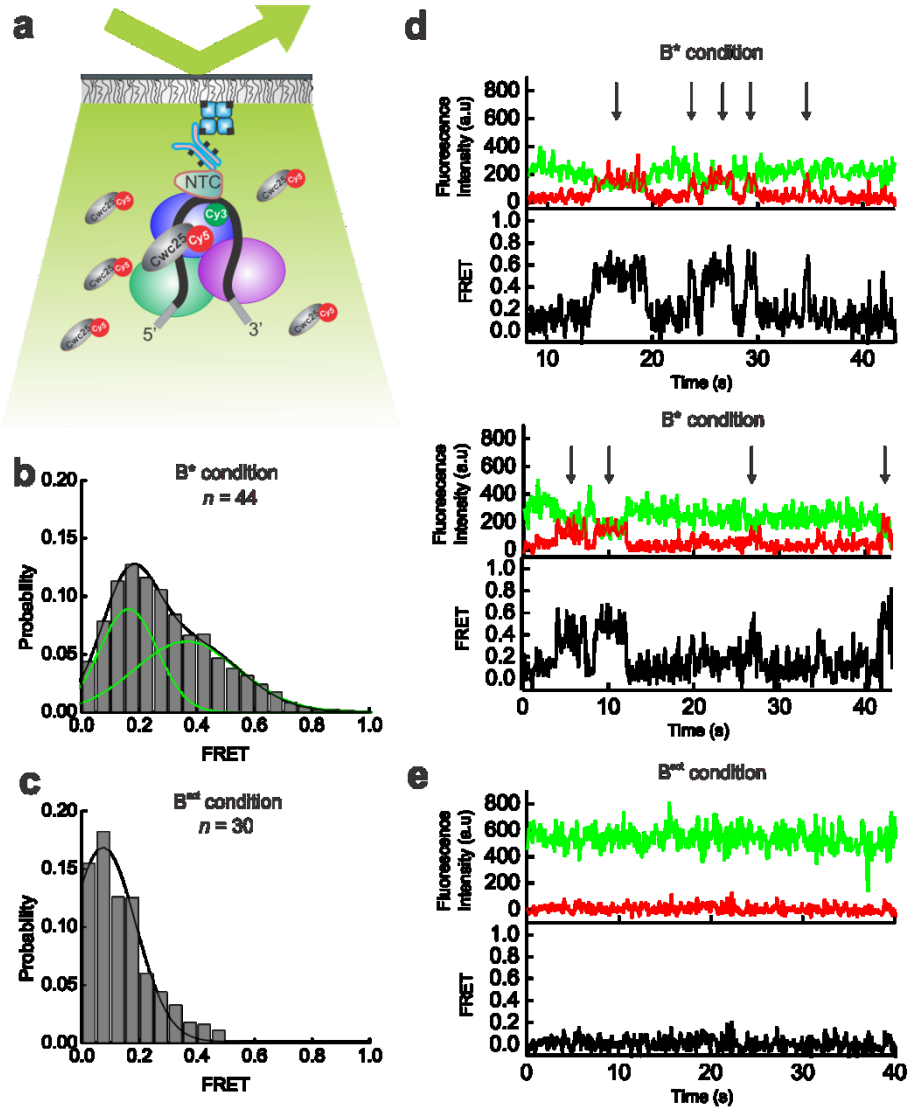



**Figure 2.11 Prp2-mediated spliceosome remodeling creates a binding site for Cwc25 near the BP**

(a) Schematic of our SiMPull experiment to test for binding of Cy5-tagged Cwc25 near the BP of spliceosome-associated pre-mRNA containing a single active Cy3 fluorophore. (b) FRET probability distribution of the raw single molecule FRET trajectories under B* conditions (in the presence of Prp2-mediated remodeling). (c) FRET probability distribution of the raw single molecule FRET trajectories under $B^{act}$ conditions (in the absence of Prp2-mediated remodeling). (d) Representative single molecule FRET trajectories showing the binding and associated FRET between the Cy5 on Cwc25 and the Cy3 near the pre-mRNA BP under $B^{*}$ conditions with the raw donor (Cy3, green), acceptor (Cy5, red), and FRET (black) trajectories. Arrows indicate binding events with close proximity to the BP. (e) Representative single molecule FRET trajectory showing the absence of FRET between the Cy5 on Cwc25 and the Cy3 near the pre-mRNA BP under $B^{act}$ conditions with the raw donor (Cy3, green), acceptor (Cy5, red), and FRET (black) trajectories.

and ATP (**Figure 2.11d,e**, $B^{act}$ condition). We conclude that Cwc25 activates the spliceosome for the first step by dynamically binding to the pre-mRNA near the BP.

## 2.4 Discussion

Here we have combined single molecule FRET between fluorophores attached near the 5'SS and BP of the pre-mRNA substrate with affinity purification, in a technique we term SiMPull[70]-FRET, to study the spliceosomal $B^{act}$ complex stalled by heat inactivation of Prp2. Stepwise addition of ATP and the recombinant proteins Prp2, Spp2 and Cwc25 revealed the role of each factor in pre-mRNA remodeling (**Figure 2.12**). We find that the pre-mRNA remains in the static low-FRET L2 state of the $B^{act}$ complex (which, for clarity, we term $L2^{act}$) with distal 5'SS and BP until the activation by Prp2–Spp2 in the presence of ATP produces the B* complex. Prp2-mediated hydrolysis of ATP (or UTP) in this step weakens the binding of some seven proteins, including SF3a and SF3b[38-40,66], that bind the pre-mRNA upstream and downstream of the BP adenosine, presumably preventing its premature nucleophilic attack on the 5'SS. Accordingly, the B*-associated low-FRET state (L2*) allows the pre-mRNA to transiently and reversibly visit two new states of either mid- (M*) or high-FRET (H*) and more proximal 5'SS and BP. First-step splicing now proceeds with low efficiency, leading to post-catalytic C complex formation signified by a static high-FRET state ($H^C$). This finding indicates that the increased proximity of the reactive sites is sufficient for catalysis. However, reaction chemistry is greatly enhanced by the addition of Cwc25, which binds the pre-mRNA substrate near the BP and slows particularly the rate constant of the high- to mid-FRET transition, leading to a longer dwell time in the pre-catalytic, stabilized FRET state $H^{C\text{-}pre}$. In turn, this event leads to enhanced progression to the static high-FRET state associated with the post-catalytic C complex ($H^C$) (**Figure 2.12**).

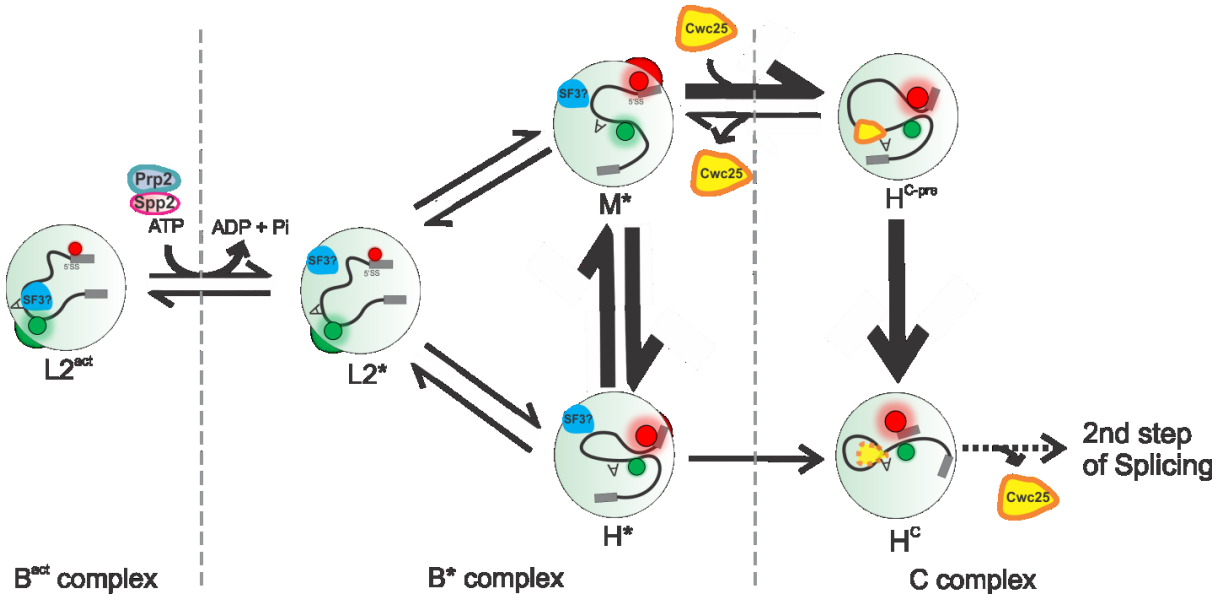



**Figure 2.12 Model for the conformational mechanism of first-step splicing.**
The 5'SS and BP of the pre-mRNA in the $B^{act}$ complex reside predominantly in the static distal $L2^{act}$ conformation with low FRET (i.e., high Cy3 and low Cy5 fluorescence, indicated by the green and red circles, respectively). Upon ATP hydrolysis and conversion into the B* complex, Prp2 along with its cofactor Spp2 unlocks the B*-associated low-FRET state L2* to reversibly sample the mid- and high-FRET (spatially proximal) conformations M* and H* (and $H^{C\text{-}pre}$ under C complex conditions). Cwc25 binds near the BP of the pre-mRNA, thus reducing the rate constant of the high- to mid-FRET transition and enhancing first-step chemistry, upon which the pre-mRNA adopts the static high-FRET state $H^{C}$.

Our data show that before the action of Prp2, Spp2, and ATP, the spliceosome keeps the reactive sites of the pre-mRNA strictly apart. This observation is consistent with and refines a recent report suggesting that stable splice-site juxtaposition occurs at some point after the NTC assembles on the pre-mRNA[54]. Furthermore, it has previously been speculated that the catalytically activated B* complex may shift back and forth between inactive and active conformations[38]. We have presented direct evidence for this hypothesis by showing that only in the B* state are dynamic excursions between low- and high-FRET states observable, and only the high-FRET state places the reactive 5'SS and BP in close enough proximity for subsequent catalysis, correlated with the appearance of the static high-FRET state $H^{C}$. The same authors also proposed[38] that Cwc25 binding may shift the equilibrium between inactive and active conformations towards the latter, which we directly observe and assign to a marked increase of the dwell time in the active conformation $H^{C\text{-}pre}$ with proximal 5'SS and BP.

The behavior of the spliceosome resembles that of a classical biased Brownian ratchet machine that draws path directionality from the random thermal fluctuations, which it constantly experiences, through a form of directional 'rectification' or 'biasing'[99,100] (**Figure 2.13**). In fact, the ribosome has been described as a biased Brownian ratchet machine[101-103] and our previous single molecule FRET probing of pre-mRNA dynamics in whole yeast cell extract suggested that the spliceosome, like the ribosome, works close to thermal equilibrium[53]. We therefore propose that the ATP-driven helicase activity of Prp2–Spp2 acts to remove SF3a–SF3b as an impediment to the intrinsic thermal fluctuations of the spliceosome-substrate complex, while Cwc25 provides directionality to the reaction pathway by then acting as a "pawl" to stabilize the catalytically competent conformation. Perhaps the closest known analogy is found in translocation of the ribosome, where the random conformational ratcheting between the two ribosomal subunits at

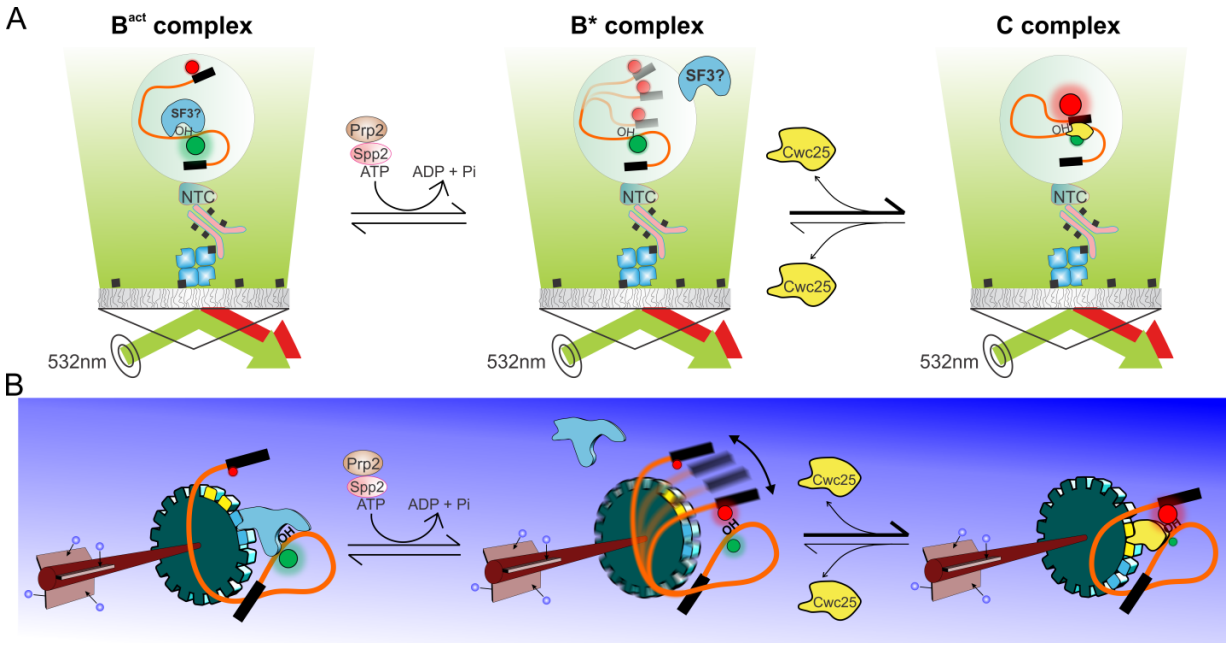



**Figure 2.13 Biased Brownian ratcheting leads to the first step of splicing**
Biochemical (A) and mechanical (B) representations of the biased Brownian ratchet mechanism utilized by the spliceosome to promote first-step splicing. Binding of the SF3a/b complex (cyan) acts as a pawl to prevent docking of the BP and 5'SS in the $B^{act}$ complex. Addition of Prp2, Spp2, and ATP results in the ATP-dependent release of SF3a/b from the spliceosome, allowing for dynamic docking and undocking of the 5'SS and BP and low levels of first-step splicing in the B* complex. Lastly, Cwc25 (yellow) acts as a new pawl, stabilizing proximal 5'SS and BP in the C complex and allowing for more efficient first-step splicing.

thermal equilibrium is rectified by GTP-bound EF-G in conjunction with the intercalation of a conserved two-nucleotide 16S ribosomal RNA "pawl" into the mRNA that appears to prevent it from ratcheting back[104]. We note that alternation of reversible thermal motion and irreversible NTP hydrolysis steps is thought to form the basis for repeated proofreading by the ribosome[105], and may do so for the spliceosome. Future studies will likely further illuminate the molecular mechanisms of these events.

Finally, DExD/H-box helicases such as Prp2 are widespread enzymes that participate in many aspects of RNA processing[106,107]. In general, they are thought to use ATP hydrolysis to remodel RNA and RNP complexes by binding, unwinding and releasing the RNA. Unlike previous single molecule approaches[53,54,108-110], our SiMPull-FRET approach has not only allowed us to unveil the dependence of pre-mRNA ratcheting on the NTPase activity of Prp2, which is then kinetically biased by Cwc25 binding, but to do so in a well-controlled purified system. Biased Brownian ratcheting may be widespread among helicase-driven RNPs, and SiMPull-FRET will allow us to test this hypothesis further.

### 2.5 Acknowledgements

We would like to thank Dr. Reinhard Lührmann (Max Planck Institute for Biophysical Chemistry, Göttingen, Germany) and R. J. Lin (Fujian Medical University, Fuzhou, China) for kindly providing plasmids for expression of Prp2, Spp2, and Cwc25; H. Hadjivassiliou and A. Price (University of California, San Francisco) for providing Cy5-body labeled actin pre-mRNA substrate; D. Semlow and J. P. Staley (University of Chicago, Illinois) for providing the dominant negative Prp16 protein.

# CHAPTER 3: Single-Molecule Cluster Analysis Identifies Signature Dynamic Conformations along the Splicing Pathway[2]

## 3.1 Introduction

Conformational dynamics play a key role in every aspect of RNA biology, such as in RNA transcription, splicing and translation[111-113]. The quantitative measurement and interpretation of these dynamics are of great importance for an understanding of the common principles underlying the biological function of RNA[112-114]. Single molecule fluorescence approaches have recently emerged as a powerful toolset to dissect the structural dynamics that form the foundation of biomolecular machines functioning at the nanometer scale[49,53,55,73,115]. For example, single molecule fluorescence energy transfer (smFRET) has been implemented to dissect spliceosome dynamics[53-55]. The spliceosome is a multi-megadalton ribonucleoprotein (RNP) complex essential for the faithful removal of introns from eukaryotic precursor messenger RNAs (pre-mRNAs) during the two chemical steps of splicing (**Figure 3.1a**)[64]. The architectural reorganization of the pre-mRNA substrate required to accommodate these two catalytic steps in a single active site are thought to be accompanied by substantial rearrangements that ensure substrate proofreading[42,59,65,116]. To explore these rearrangements, we have labeled the efficiently splicing yeast pre-mRNA Ubc4[53,117] with the FRET pair Cy5 and Cy3 seven nucleotides

[2] *Nature Methods*, in revision. Matthew Kahlscheuer performed most of the smFRET experiments and the corresponding data analysis. Mario Blanco performed a few of the smFRET experiments and the corresponding data analysis, as well as helped with the development of the SiMCAn software. Joshua Martin wrote the MATLAB scripts used in the SiMCAn software.

upstream of the 5' splice site (5'SS) and six nucleotides downstream of the branch point (BP), respectively. This approach yields a substrate capable of detecting changes in intron conformation as a result of 5'SS and BP (un)docking (**Figure 3.1a,b**) that we previously used to show that one of several DExD/H-box ATPases, Prp2, unlocks intrinsic conformational dynamics in the isolated spliceosomal $B^{act}$ complex, setting the stage for first-step catalysis through a biased Brownian ratcheting mechanism[55].

Despite years of utilization, the quantitative methods available for an in-depth dissection of the dynamics observed in smFRET studies are still limited. In particular, the sheer complexity of the dynamics encountered in many molecular machines, such as the spliceosome, with often a large number of conformations, only limited and transient enrichment of any one species, and mostly asynchronous and often heterogeneous kinetics, render the current state-of-the-art analysis of individual state transitions as independent stochastic events insufficient for an in-depth understanding of the mechanisms of action underlying biological function. To extract additional information, several recent studies have analyzed common smFRET metrics more thoroughly, specifically FRET probability histograms and state-to-state transition kinetics (**Figure 3.1c,d**)[73]. For example, it has been theoretically demonstrated that in certain favorable cases interstate dynamics can be extracted from histograms through an analysis of photon arrival times and lifetimes[118], requiring sophisticated and less common pulsed-laser instrumentation. In addition, state-to-state transition kinetics have been extracted utilizing clustering algorithms to identify distinct kinetic behaviors[119,120]. All of these approaches have focused on small datasets with 2-3 FRET states and limited dynamics. Unfortunately, they are limited when more complex systems with multiple states and complex kinetic networks are examined under non-equilibrium conditions.

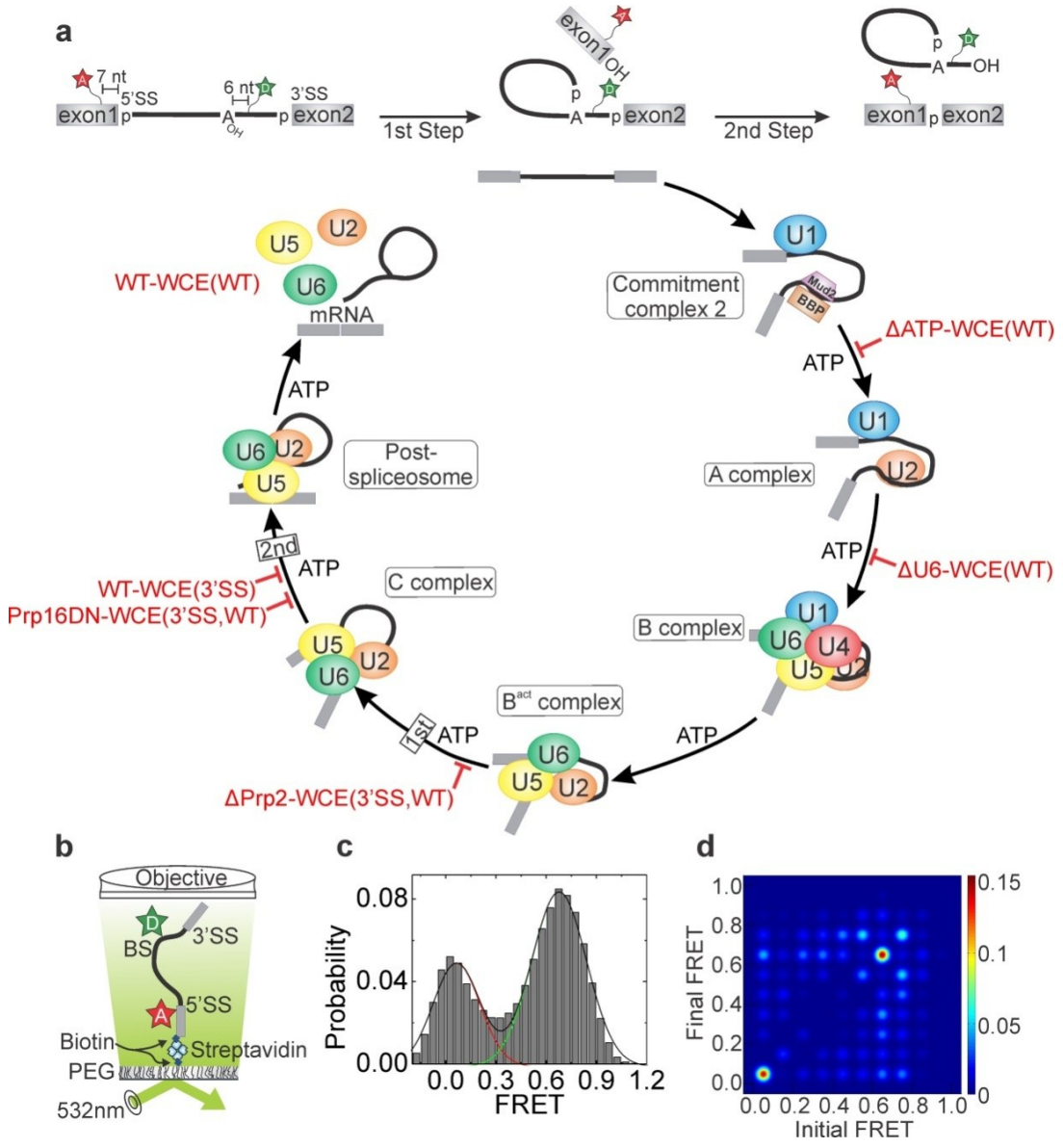



**Figure 3.1 Single molecule fluorescence energy transfer (smFRET) of pre-mRNA splicing**

(a) The fluorescent substrate used to monitor pre-mRNA dynamics contains Cy5 and Cy3 fluorophores seven nucleotides upstream of the 5'SS and six nucleotides downstream of the BP, respectively. The spliceosome assembly and catalysis pathway is thought to progress in a stepwise manner requiring ATP at several steps of assembly. The biochemical and genetic stalls utilized in this study are indicated by red blocks. **(b)** Prism-based TIRFM setup for smFRET. **(c)** FRET probability distribution analysis confirms diverse ensemble behaviors in the pre-mRNA conformation at various stages of spliceosome assembly; as example data from the ΔPrp2-WCE(3'SS) condition are shown. **(d)** Transition Probability Density Plots (TODPs) highlight the fraction of molecules that remain static (diagonal) or transition between two FRET states (off-diagonal); as example data from the ΔPrp2-WCE(3'SS) condition are shown.

We present here a method that utilizes hierarchical clustering as a means to group and sort smFRET trajectories and identify commonalities in a vast dataset of high complexity. We termed this tool Single Molecule Cluster Analysis (SiMCAn) and used it to characterize the pre-mRNA dynamics associated with the assembly and catalytic steps of the yeast spliceosome. Exploiting eight independent depletion conditions and mutations to block the splicing cycle at specific points, dynamic behaviors were assigned to specific complexes. SiMCAn reduces every single molecule trajectory, regardless of its number of states, to an easily comparable unit of information, the FRET Similarity Matrix (FSM). By leveraging hierarchical clustering techniques adapted from evolutionary analysis, we identified common dynamic behaviors across 10,680 different Ubc4 pre-mRNA molecules. Importantly, we accomplished an unbiased, model-free identification of commonalities and differences between splicing complexes through a second level of clustering based on the abundance of dynamic behaviors exhibited by defined functional intermediates. Applying SiMCAn to selectively stalled splicing reactions thus allowed us to efficiently assign pre-mRNA FRET states and transitions to specific splicing complexes, including a heretofore undescribed low-FRET conformation adopted late in splicing by a 3' splice site mutant. These results establish SiMCAn as an effective bioinformatics tool to characterize complex smFRET behavior of dynamic cellular machines.

## 3.2 Materials and Methods

### 3.2.1 Synthesis of pre-mRNA substrates

The Ubc4 pre-mRNA substrates used in this study (**Table 3.1**) were synthesized as previously described[53]. Briefly, the 135-nucleotide pre-mRNA was ligated from two fragments: a 59-nucleotide 3' segment with 5-amino-allyl-uridine at the +6 position relative to the BP adenosine and a 76-nucleotide 5' segment with 5-amino-allyl-uridine at the -7 position

| | |
|---|---|
| Ubc4 Wildtype (WT) | 5'-biotin-GAACUAAGUGAUC**(5-N-U)**AGAAAG*GUAUGUCUAAAGU UAUGGCCACGUUUCAAAUGCGUGCUUUUUUUUUAAAACU UAUGCUCUUAUUUACUA**A**CAAAA**(5-N-U)**CAACAUGCUAUUG AACUA**G**A**G**AUCCACCUACUUCAUGUU-3'* |
| Ubc4 3' Splice Site (3'SS) | 5'-biotin-GAACUAAGUGAUC**(5-N-U)**AGAAAG*GUAUGUCUAAAGU UAUGGCCACGUUUCAAAUGCGUGCUUUUUUUUUAAAACU UAUGCUCUUAUUUACUA**A**CAAAA**(5-N-U)**CAACAUGCUAUUG AACUA**C**A**C**AUCCACCUACUUCAUGUU-3'* |
| DNA splint | 5'-GTTGATTTTGTTAGTAAATAAG(SP9)GTTTTAAAAAAAAAGCACGC-3' |
| D1 Oligo | 5'-ATCTCTGTATTGTTTCAAATTGACCAA-3' |

**Table 3.1 Sequence information of the oligonucleotides used in this study**
The Ubc4 intron is italicized, and the BP adenosine is bold and underlined. The red and green "(5-N-U)" denote the allyl-amine modified uridines used to attach the Cy5 and Cy3 fluorophores. In the 3'SS mutant, the two bold and underlined cytosines replace guanines in the wild-type sequence. The DNA splint is the oligonucleotide used for templated ligation during synthesis of the WT and 3'SS pre-mRNA substrates. Sp9 denotes a 9-carbon linker.

relative to the 5'SS. The 3'SS mutant had the guanines at positions 115 and 117 on the 3' segment replaced with cytosines. The 5' and 3' fragments were coupled to Cy5 and Cy3 N-hydroxysuccinimidyl ester (GE Healthcare), respectively, by resuspending 4 nanomoles of RNA in 40 µl of 0.1 M sodium bicarbonate buffer, pH 9.0, and incubating for 30 min at 60 °C with the proper dye pack dissolved in DMSO. The conjugated fragments were ethanol precipitated and washed with 70% (v/v) ethanol to remove unconjugated dye. Unlabeled RNA was removed by purification on benzoylated naphthoylated DEAE (BND)-cellulose (Sigma) that was washed with 1 M NaCl containing 5% (v/v) ethanol. Fully labeled RNA fragments were eluted with 1.5 M NaCl containing 20% (v/v) ethanol and further precipitated to remove excess salt. Labeled fragments were combined with an equal molar amount of DNA splint (**Table 3.1**) and ligated by incubating with RNA Ligase 1 (NEB) for 4 h at 37 °C as described[53,117]. Full length, labeled Ubc4 was then purified on a denaturing 7 M urea, 15% (w/v) polyacrylamide gel.

3.2.2 Preparation of yeast whole cell extract

Splicing active whole cell extract (WCE) was prepared from either yeast strain BJ2168 or a *prp2-1 cef1-TAP* yeast strain (ATCC 201388: *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0)* as previously described[53,121]. Briefly, cells were grown in YPD medium to an OD600 of 1.6-2.0 before they were harvested and washed in AGK buffer (10 mM HEPES-KOH, pH 7.9, 1.5 mM $MgCl_2$, 200 mM KCl, 10% (v/v) glycerol, 0.5 mM DTT, 0.6 mM PMSF, and 1.5 mM benzamidine). A thick slurry of cells was dripped into liquid nitrogen to form small cell pellets that could be stored at -80 °C. The frozen pellets were disrupted by manual grinding with a mortar and pestle half-submerged in liquid nitrogen for 30 min. The resulting frozen powder was thawed in an ice bath and centrifuged at 17,000 rpm in a type 45 Ti Beckman rotor. The supernatant was then centrifuged at 37,000 rpm in a Ti-70 rotor for 1 h. The clear middle layer

was removed with a syringe and dialyzed for 4 h against 20 mM HEPES-KOH, pH 7.9, 0.2 mM EDTA, 0.5 mM DTT, 50 mM KCL, 20% (v/v) glycerol, 0.1 mM PMSF, and 0.25 mM benzamadine with one buffer exchange.

3.2.3 Accumulation of splicing complexes

**Table 3.2** describes all experimental conditions by identifying the substrate and WCE used along with the complex formed. All splicing products were confirmed via *in vitro* splicing assays by incubating 4 nM fluorescent Ubc4 in splicing buffer (8 mM HEPES-KOH, pH 7.0, 2 mM $MgCl_2$, 0.08 mM EDTA, 60 mM $K_i$(PO4), 20 mM KCl, 8% (v/v) glycerol, 3% (w/v) PEG, 0.5 mM DTT) and 40% (v/v) WCE at 25 °C for 40 min. Products were analyzed by separation on a 7 M urea, 15% (w/v) polyacrylamide gel and scanned on a Typhoon variable mode imager (GE Healthcare, **Figure 3.2**). ATP depletion was performed by pre-incubating WCE with 1 mM glucose at 25 °C for 10 min prior to incubation with splicing buffer and substrate. Endogenous U6 snRNA was depleted by pre-incubation of WCE with 300 nM D1 oligodeoxynucleotide (**Table 3.1**) in splicing buffer, 50% (v/v) WCE, and 2 mM ATP at 33 °C for 30 min prior to incubation with substrate. Knockdown of endogenous Prp2 was performed by heating *prp2-1 cef1-TAP* WCE to 37 °C for 40 min prior to incubation with splicing buffer, ATP, and pre-mRNA substrate. Endogenous Prp16 was inactivated using 100 nM of a Prp16 dominant-negative mutant (Prp16DN; K379A) added to the BJ2168 WCE for 10 min prior to incubation with splicing buffer, 2mM ATP, and pre-mRNA substrate. On-slide splicing assays were performed as the *in vitro* splicing assays with the exception that all materials were combined prior to flowing reaction mixtures onto a substrate-coated, PEG-passivated slide using established procedures[53,55].

| Condition | Splicing Complex | Substrate | Extract |
|---|---|---|---|
| ΔATP-WCE(WT) | CC2 | WT | BJ2168 extract + 1mM glucose |
| ΔU6-WCE(WT) | A complex | WT | BJ2168 extract + 300nM D1 |
| ΔPrp2-WCE(WT) | $B^{act}$ Complex | WT | *Prp2-1 Cef1-TAP* strain extract |
| ΔPrp2-WCE(3'SS) | $B^{act}$ Complex | 3'SS | *Prp2-1 Cef1-TAP* strain extract |
| Prp16DN-WCE(WT) | C Complex | WT | BJ2168 extract + Prp16DN mutant protein |
| Prp16DN-WCE(3'SS) | C Complex | 3'SS | BJ2168 extract + Prp16DN mutant protein |
| WT-WCE(3'SS) | C Complex | 3'SS | BJ2168 extract |
| WT-WCE(WT) | Post-spliceosome | WT | BJ2168 extract |

**Table 3.2 Substrate and extract used to form each of the splicing complexes**

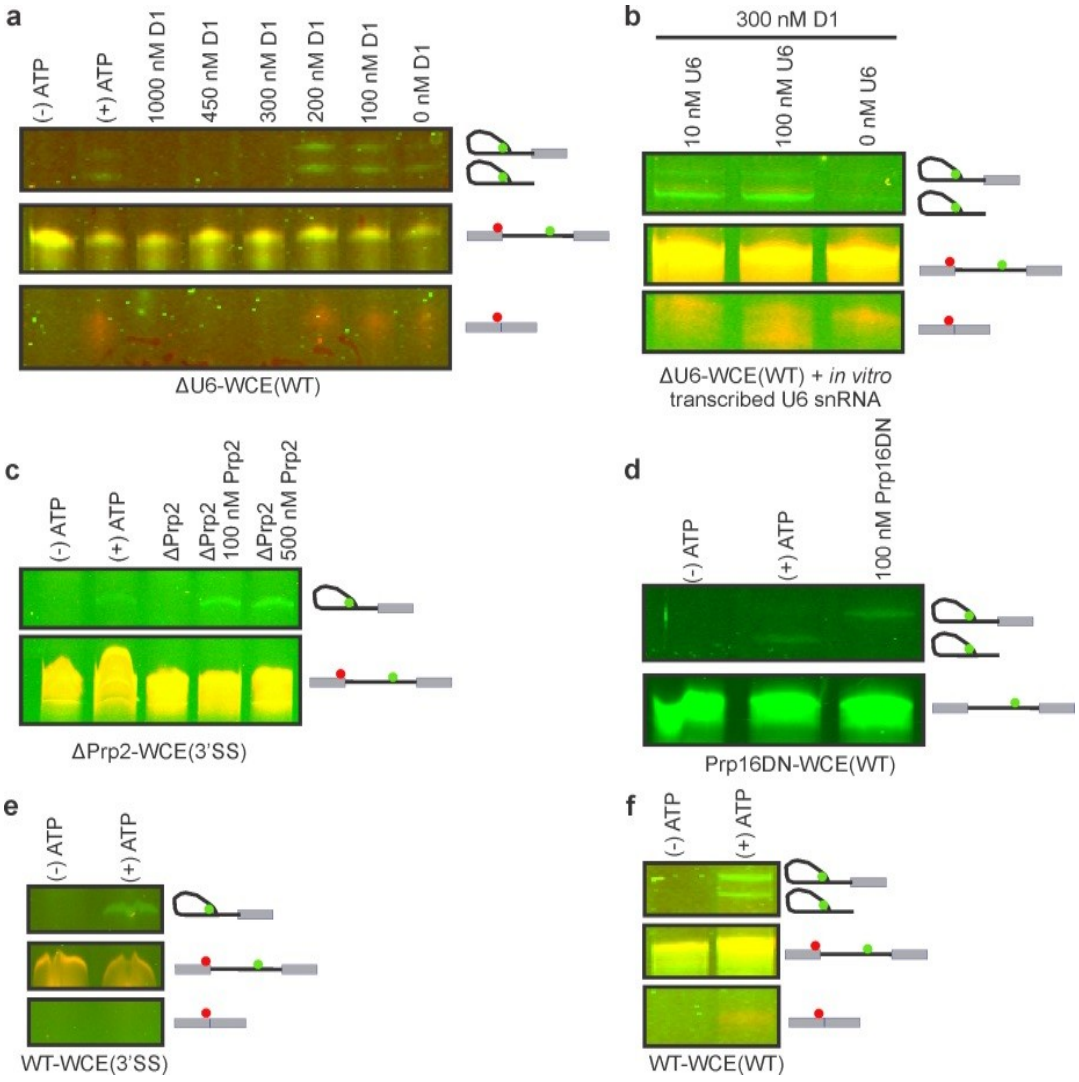



**Figure 3.2 Confirmation of blockage and reconstitution of splicing by *in vitro* splicing assays**

Denaturing, 7 M urea, 15% (w/v) polyacrylamide gels were scanned with a variable mode Typhoon imager. The intron and intron-lariat products are observed in the Cy3 scan (green) and the mature mRNA product is visualized in the Cy5 scan (red). **(a)** The optimized concentration of D1 required to deplete U6 snRNA (300 nM) was determined by titrating increasing amounts of the oligodeoxynucleotide into the *in vitro* splicing assay. **(b)** Using the previously determined optimal concentration of D1 (300 nM, **a**), extract viability was confirmed through reconstitution with *in vitro* transcribed U6 snRNA. **(c)** Incubation of extract at 37 °C for 40 min completely blocks splicing activity (ΔPrp2 lane). Addition of recombinant Prp2p to ΔPrp2 extract results in reconstitution of splicing, as expected. **(d)** Addition of recombinant dominant mutant Prp16DN to yeast extract stalls splicing after the first chemical step. **(e)** Incubation of WT-WCE with 3'SS mutant substrate stalls splicing after the first step while incubation with a WT substrate **(f)** results in efficient progression through both steps of splicing.

### 3.2.4 Single molecule FRET

Single Molecule FRET was carried out in the same manner as previously described[53,55]. Using a prism-based TIRF microscope[45,49,122], we collected data from single molecules incubated under the desired conditions (**Table 3.2**). Data were collected from two to three fields of view for each time period of 0-8 min (early), 18-23 min (middle), and 33-40 min (late) after addition of WCE. The donor (Cy3) and acceptor (Cy5) fluorophores were excited using a 532- and 635-nm laser, respectively, with the resulting emission recorded at 100 ms time resolution with a Princeton Instruments, I-PentaMAX intensified CCD camera. A FRET value was calculated by dividing the intensity of the acceptor emission by the total emission from both donor and acceptor.

### 3.2.5 Single molecule cluster analysis – SiMCAn

Each individual FRET trace was fitted with individual Hidden Markov Models (HMMs) of up to 10 states using vbFRET[123] in Mathwork's MATLAB environment with no assumptions about the values or distributions. The resulting paths were then assigned to the closest of 10 evenly spaced states (0.05-0.95, increment of 0.10 as our resolution limit). Traces of less than 3 s (30 frames) length were discarded and a transition probability (TP) matrix was constructed for each of the remaining molecule traces. Each TP matrix was then combined with the vector describing the percent of the trace that occupies each FRET state to create a FRET similarity matrix (FSM). The FSMs were divided into categories containing static traces and dynamic traces, the dynamic traces identified and characterized by having at least one FRET transition between two FRET states. Static traces were identified automatically based on their unique signature with just a single FRET value and kept separate for the remaining analysis. Static molecules could arise due to fluorophores photobleaching prior to a transition taking place. On the other hand, formation of a particular complex may lead to a very stable, unchanging conformation that results in emission

of a single (static) FRET state. Dynamic traces were used as input for a hierarchical clustering analysis performed by MATLAB. The resulting hierarchical tree was then used to identify clusters of traces with similar behavior as identified from their FSM. The tree was pruned at a height that resulted in 25 dynamic clusters in addition to 10 static clusters as assigned by their FRET state. The height used to determine the clusters in the hierarchical tree was determined using an iterative measurement of the inter-cluster distances and a modified k-means algorithm. The specific cut-off was chosen at the first point where randomly assigned traces had a higher inter-cluster distance than the hierarchical clustering. The resulting clusters were analyzed and labeled according to their occupancy in the FRET states. All analysis and descriptions of the clusters were performed using MATLAB.

### 3.2.6 Generation of simulated dataset

Artificial HMMs containing the distinctions of interest were used to generate traces of $10^6$ time step length. These traces were used to generate 1,500 subtraces where the starting points were uniformly selected along the full trace and the length determined by a Poisson distribution with a lambda of 100. The resulting traces were treated exactly like experimentally acquired data for analysis by SiMCAn.

## 3.3 Results

### 3.3.1 Hierarchical clustering of complex smFRET behaviors

State-to-state transitions in single molecule trajectories report on the accessibility of conformational states and their ability to interconvert. Hidden Markov Models (HMMs) are the most commonly utilized tools for identifying state-to-state transitions. To identify the commonalities and differences across hundreds or thousands of smFRET traces, SiMCAn

generates a FRET similarity matrix (FSM) for each trajectory by fitting it with a HMM to reliably identify the number of FRET states and their FRET values (**Figure 3.3a**). It should be noted that, although vbFRET[123] was utilized for HMM analysis of our dataset, any HMM fitting tool can be utilized that satisfies the user's fitting preferences. To enable a direct comparison across a large dataset, we binned each FRET state into one of ten evenly spaced FRET values (0.05-0.95, with increments of 0.10) (**Figure 3.3b**) that together evenly span the viable FRET range and are commensurate with typical signal-to-noise ratios. The resulting HMMs are used to construct transition probability (TP) matrices between states (**Figure 3.3c**). Each TP matrix is then combined with the occupancies of the individual FRET states to create the final FSM. Prior to clustering analysis, molecules with no transitions (static) are automatically identified and analyzed separately. The remaining dynamic molecule FSMs are used as input for hierarchical clustering analysis, an agglomerative clustering technique that aims to group data of similar characteristics[124,125]. The distance between FSMs is found using the Euclidean distance between any two matrices (Methods). The result of this clustering is a hierarchical tree, where each leaf on the tree represents the dynamics of an individual molecule. Branch points in the hierarchy indicate a split in dynamic behavior of the group of molecules at a given level of coarseness (**Figure 3.3d**). Throughout, each cluster is represented using the average TP matrix (**Figure 3.3e**), a collection of traces (**Figure 3.3f**), and the probability distribution of FRET states within the cluster (**Figure 3.3g**). This model-free clustering of a combined large smFRET dataset is designed to enable an unbiased and quantitative classification of single-molecule dynamic behavior throughout the entire splicing pathway.

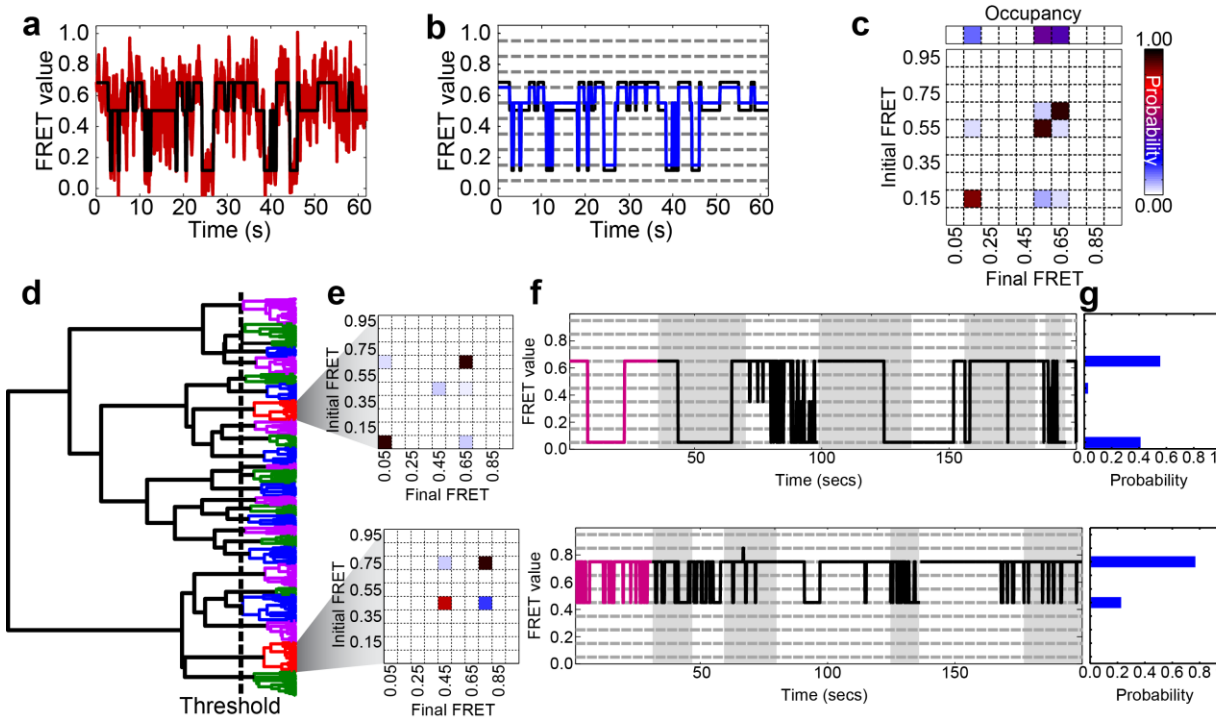



**Figure 3.3 Single molecule cluster analysis (SiMCAn) sorts and clusters molecules that share common dynamic behaviors**

**(a)** Representative raw (red) and corresponding idealized FRET trace fit using the idealizing hidden Markov model (HMM, black). **(b)** Idealized FRET trace (black) and reassigned FRET trace (blue) to the closest of 10 evenly spaced states (0.05-0.95, increment of 0.10). **(c)** Transition probability (TP) matrix describing the transitions for the molecule in **a**. **(d)** Hierarchical tree as a result of hierarchical clustering analysis using all 6,079 dynamic molecules. Each colored branch describes a set of molecules that shares common FRET transition probabilities. The dashed line indicates the threshold of 25 clusters used to describe the data. **(e to g)** Cluster description for 2 of the 25 dynamic clusters of the full splicing dataset. Each representation shows the TP matrix of the cluster **(e)**, the trace closest to the cluster center (magenta) and up to 200 s of random (black) traces from the cluster **(f)**, and the probability of FRET states within the cluster **(g)**. Grey and white backgrounds demarcate individual trajectories in **(f).**

3.3.2 Validation of SiMCAn using simulated datasets

To evaluate whether SiMCAn is able to correctly identify and segregate trajectories with known FRET states, we applied it first to a simulated dataset containing 1,500 trajectories that reversibly transition from a 0.15 to a 0.45 FRET state and an equal number of trajectories that transition from the same 0.15 FRET state to a 0.85 state instead (**Figure 3.4a**), with average rate constants of $k_{0.15\rightarrow0.45}$ = 0.54 $s^{-1}$, $k_{0.45\rightarrow0.15}$ = 0.53 $s^{-1}$ and $k_{0.15\rightarrow0.85}$ = 0.53 $s^{-1}$, $k_{0.85\rightarrow0.15}$ = 0.53 $s^{-1}$, respectively. SiMCAn properly identified and separated these two molecular behaviors (**Figure 3.4b)**, demonstrating that trajectories can easily be clustered based on the identity of their FRET states. A second and more important feature of SiMCAn is the ability to segregate trajectories based on differing kinetics. We analyzed a second set of 3,000 simulated trajectories possessing two FRET states of 0.15 and 0.75, with half of the trajectories designed to have identical interconversion rate constants of 0.54 $s^{-1}$ whereas the other half transition much more slowly with rate constants of 0.15 $s^{-1}$ (**Figure 3.4c)**. SiMCAn identified two clusters with distinct transition rate constants between the 2 states (**Figure 3.4d**). These results demonstrate SiMCAn's ability to differentiate FRET trajectories based on their FRET states and kinetics.

3.3.3 Validation of SiMCAn using purified spliceosomal complexes

To benchmark SiMCAn against an experimental dataset with inherent limitations imposed by, e.g., signal noise and premature photobleaching, we chose to analyze a previously published dataset collected during the Prp2-mediated conformational transition immediately prior to the first step of splicing[55]. Briefly, the immobilized $B^{act}$ complex containing FRET-labeled Ubc4 was monitored as it progresses through the B* to the C complex upon addition of recombinant proteins Prp2, Spp2 and Cwc25 (**Figure 3.5a**). Our previous FRET probability analysis indicated a dramatic shift from a 0.3- or low-FRET state in $B^{act}$ to a 0.7- or high-FRET state in the C

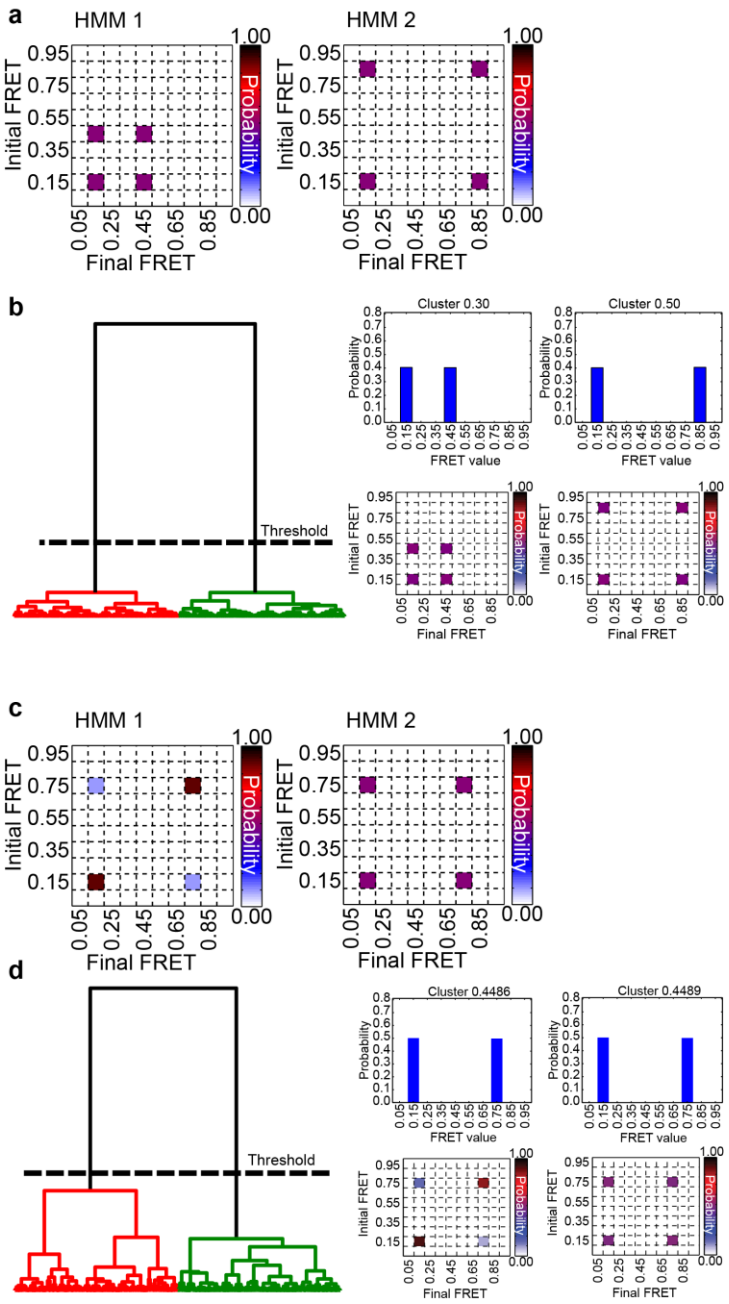



**Figure 3.4 Clustering of simulated datasets**
**(a)** Transition probability (TP) matrices possessing one shared FRET state (0.15) and one differing FRET state (0.45 or 0.85) that were used to generate the 1500 random traces for clustering by SiMCAn. **(b)** Hierarchical tree showing the two cluster threshold found by SiMCAn and the two resulting cluster probability histograms and TP matrices. **(c)** TP matrices possessing the same two FRET states but different rates of interconversion used to generate the 1500 random traces for clustering by SiMCAn. **(d)** Hierarchical tree showing the three cluster threshold found by SiMCAn and the three resulting cluster probability histograms and TP matrices.

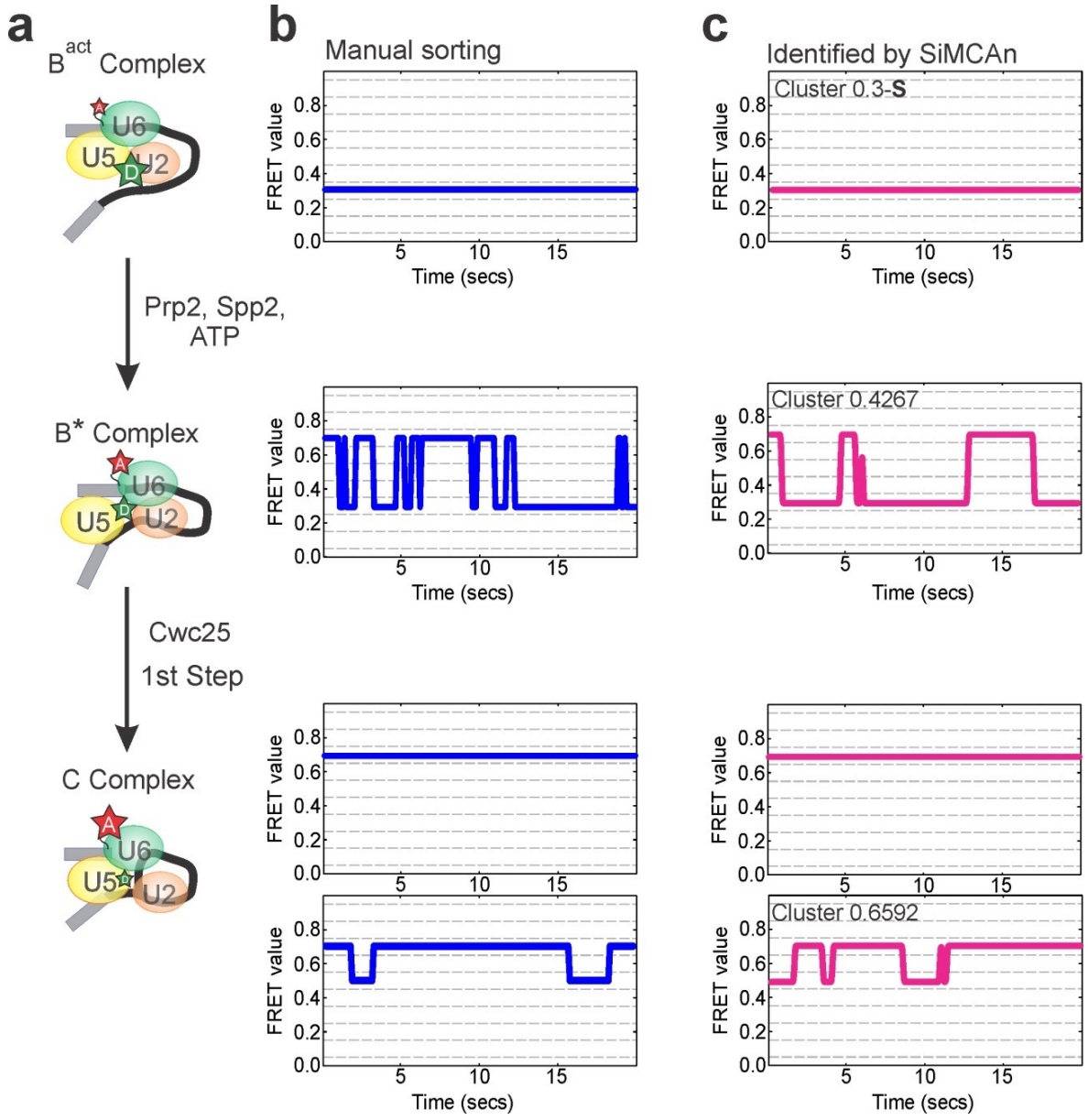



**Figure 3.5 Validation of SiMCAn using a previously analyzed dataset describing the transition from the purified $B^{act}$ to the C complex**

(a) Protein requirements for the transition from the $B^{act}$ complex through B* to the C complex. (b) Visually most representative FRET trajectories describing the dominant behavior of molecules in each complex found through manual sorting of traces. (c) The smFRET trajectories found using SiMCAn that are most similar to the cluster center of the four highlighted clusters (**Figure 3.6c**) that describe each of the splicing complexes. Dynamic clusters are labeled by the weighted average FRET value of the molecules within the cluster (e.g., 0.2563) while static clusters are labeled by the single state they describe (e.g., 0.1-**S**).

complex (**Figure 3.6b**), but revealed little about the underlying mechanism. Only upon months of manually sorting molecules did we find that the transition from the pre-catalytic $B^{act}$ complex through the post-catalytic (C) complex requires, first, the ATPase activity of the DExD/H-box ATPase Prp2 to unlock the pre-mRNA from a static low-FRET state (**Figure 3.5b**), allowing it to make dynamic excursions into transient high-FRET conformations associated with the intermediate (B*) complex (**Figure 3.5b**). Second, we found that Cwc25 is required to enrich C complex formation as evident from a stabilized high-FRET state, indicative of first-step splicing (**Figure 3.5b**)[55]. In addition to this static high-FRET population, we observed a significant fraction of dynamic molecules transitioning between a long-lived high-FRET state and a shorter-lived, 0.5- or mid-FRET state (**Figure 3.5b**).

Notably, SiMCAn was able to rapidly (within minutes) and correctly identify these previously only manually identified[55] sub-populations of pre-mRNA molecules. To this end, the HMM-fitted FRET traces under $B^{act}$, B*, and C complex conditions were combined and analyzed using SiMCAn to determine if the analysis could recapitulate the manual annotation of these traces. Maximizing the inter-cluster distances while minimizing the intra-cluster distances using SiMCAn revealed 9 dynamic and 4 static clusters as best fitting the data (**Figure 3.7**). These clusters were combined into a single bar graph to depict the fraction of molecules that occupy each cluster under each experimental condition (**Figure 3.6c**). Reproducing our previous analysis, a cluster of molecules adopting a static low-FRET state (0.3-**S**) was identified as dominant under $B^{act}$ conditions (**Figure 3.5c**), whereas a static high-FRET cluster (0.7-**S**) was most abundant under C complex conditions (**Figure 3.5c**). In addition, SiMCAn identified two dynamic clusters significantly populated under B* and C complex conditions

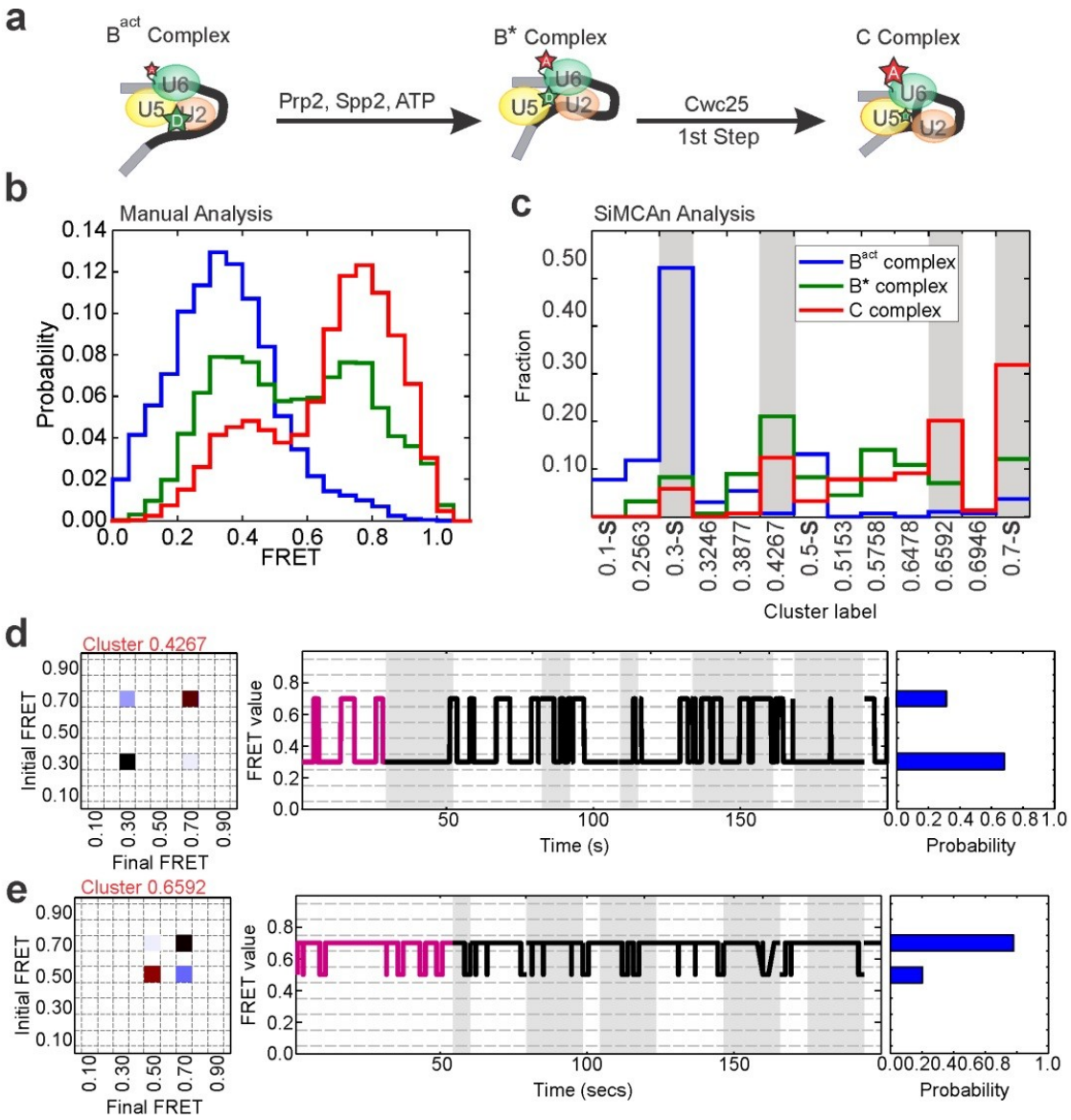



**Figure 3.6 Validation of SiMCAn using previously analyzed data**

Validation of SiMCAn using previously analyzed data describing the transition of the $B^{act}$ complex through the C complex and first step of splicing (a). (b) FRET probability analysis for molecules in the $B^{act}$ (blue), B* (green), and C complexes (red). (c) Cluster occupancy histogram showing the fraction of molecules from each experimental condition that occupy the 9 dynamic and 4 static clusters found using SiMCAn. Dynamic clusters were labeled by the weighted average FRET value of the molecules within the cluster (e.g., 0.2563) while static clusters are labeled by the single state they describe (e.g., 0.1-**S**). (d and e) Dynamic clusters enriched in the B* (cluster 0.4267) and C (cluster 0.6478) complexes. Each representation shows the TP matrix of the cluster (left), the closest (magenta) and several random (black) traces from the cluster (middle), and the probability of FRET states within the cluster (right)

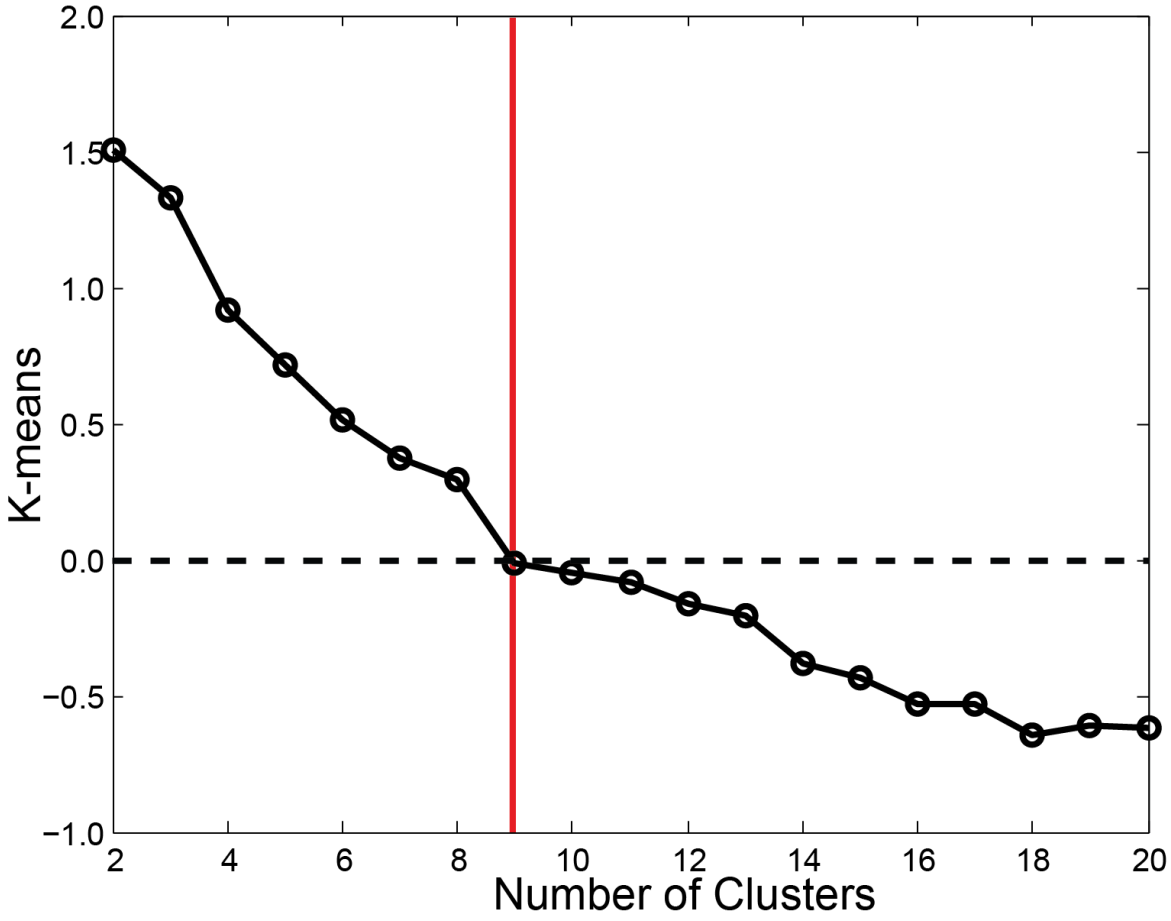



**Figure 3.7 Cluster number determination for the $B^{act}$ dataset**
Iterative measurement of inter-cluster distances using a modified k-means algorithm utilized to determine the number of clusters that best describes the previously analyzed $B^{act}$ dataset.

(**Figure 3.6c**). Cluster 0.4267 contains molecules with a short-lived high-FRET state and longer dwell times in the low-FRET state that are most abundant under B* conditions (**Figure 3.5c**). By contrast, cluster 0.6592 contains molecules with a longer-lived high-FRET state featuring rapid excursions back to a mid-FRET state that are significantly enriched upon addition of Cwc25 to form the C complex (**Figure 3.5c),** matching our previous manual analysis[55]. These results demonstrate that SiMCAn is not only able to segregate experimental molecule trajectories based on their FRET states in an unsupervised and consistent fashion, but also to classify differences in state-to-state interconversion kinetics accurately.

3.3.4 Biochemical and genetic stalls of the spliceosome lead to distinct behaviors

Having established that SiMCAn will reveal known dynamic behaviors in simulated (**Figure 3.4**) and experimental smFRET trajectories (**Figure 3.5**), we next utilized it on new single molecule trajectories enriched for specific stages of splicing through the use of biochemical and genetic stalls. smFRET data were collected upon incubation of FRET-labeled Ubc4 pre-mRNA with yeast whole cell extract (WCE), allowing for spliceosomal assembly on and splicing of the fluorescent substrate. Time courses were performed during which smFRET was recorded within time windows 0-8 min (early), 18-23 min (middle) and 33-40 min (late) after addition of WCE. To assign dynamics to particular splicing intermediates without a need for cumbersome biochemical isolation, we chose to utilize eight biochemical, yeast genetic, and substrate mutation stalls, and combinations thereof, known to allow for efficient accumulation of specific splicing intermediates and thus particular FRET behaviors in WCE (**Figure 3.1a** and **Table 3.2**). Blockage and release by reconstitution were verified by bulk *in vitro* splicing assays in yeast WCE (**Figure 3.2**). smFRET data for each stall were then acquired using the same time lapse approach utilized for the WT-WCE(WT) condition. FRET probability distributions (**Figure 3.8**)

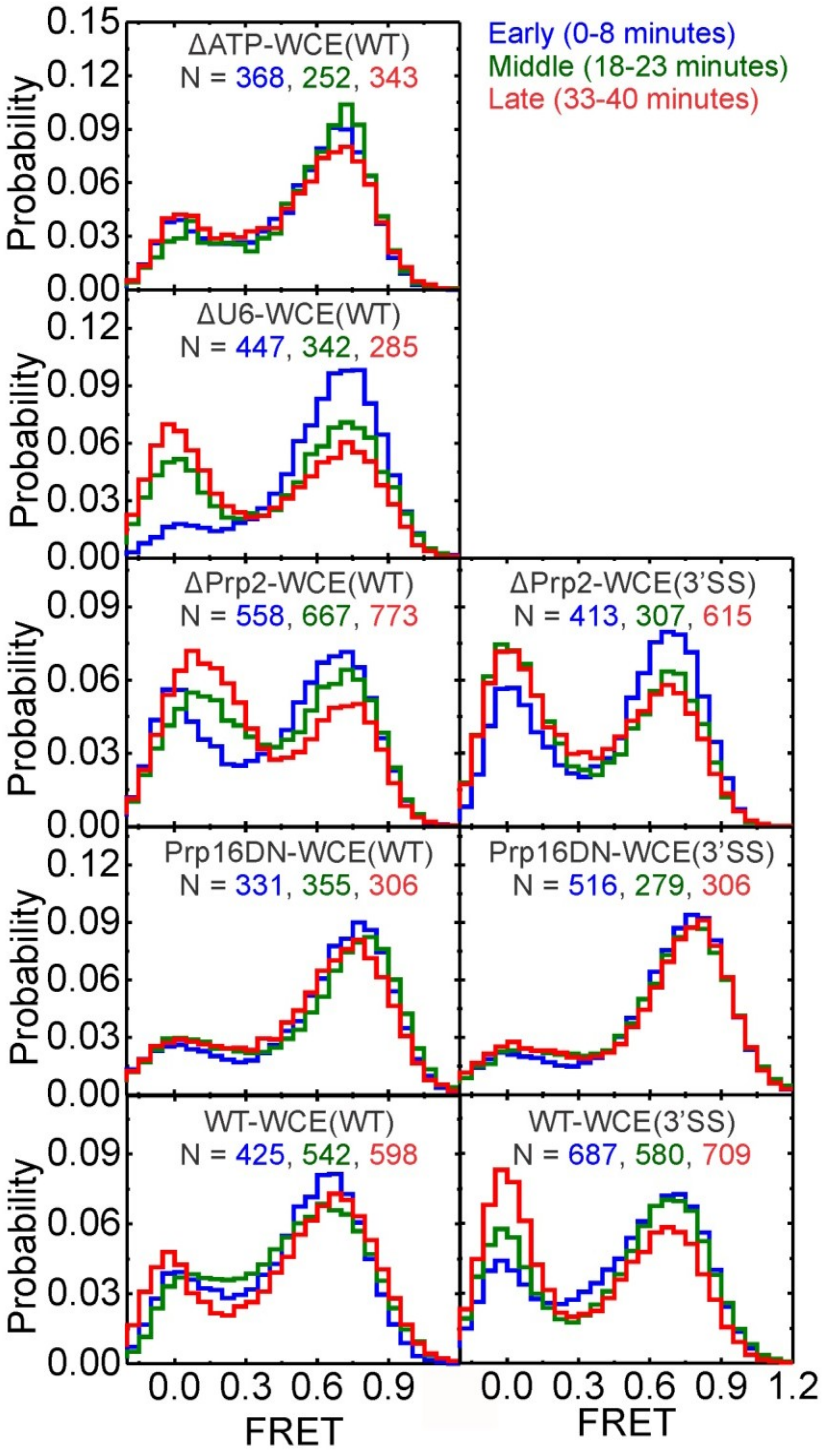



**Figure 3.8 FRET probability distribution analysis**
FRET probability distribution analysis for each of the 8 experimental conditions over the time course of the smFRET experiments.

and transition occupancy density plots (TODPs, **Figure 3.9**) were utilized to broadly summarize the behavior of hundreds of molecule trajectories per condition[73], confirming that the blocks lead to different ensemble averaged temporal behaviors. However, this far more complex data set is not amenable to standard analysis techniques as it includes an unprecedented number of traces from splicing complexes stalled by mutation throughout the splicing cycle. As such, it represents an ideal novel application for SiMCAn.

3.3.5 Identifying biologically defined dynamics using SiMCAn

Application of SiMCAn to this new comprehensive dataset allowed us to identify and cluster sets of molecules that share common dynamic behaviors. Prior to clustering, 4,601 static molecules were identified and analyzed separately. The hierarchical tree of the remaining 6,079 dynamic traces was pruned to a height that led to 25 distinct clusters (**Figure 3.3d**), determined through an iterative measurement of inter-cluster distances using a modified k-means algorithm (**Figure 3.10**)[126], so that each cluster represented a unique dynamic behavior (**Figure 3.3e-g**, **Figure 3.11**). Static clusters were named by their sole FRET state (e.g., 0.05-**S**), whereas dynamic cluster names were assigned based on the first and second most occupied FRET states within the cluster (e.g., cluster 0.65-0.05 primarily occupies 0.65 and 0.05 FRET states). To determine the robustness of SiMCAn's clustering approach, we performed bootstrapping by utilizing the transition probability matrices of four of the 25 SiMCAn-identified dynamic clusters to generate a new set of 4,500 randomized trajectories as input for clustering by SiMCAn (**Figure 3.12a**). As expected, SiMCAn identified four clusters of behaviors with FRET states and dynamic behaviors that can be directly mapped to each of the input HMMs (e.g. Cluster 0.05-0.25 maps to Cluster 0.05-0.25, **Figure 3.12b**). This analysis supports the notions that SiMCAn robustly identifies FRET states and their interconversion kinetics in increasingly complex datasets, and that the

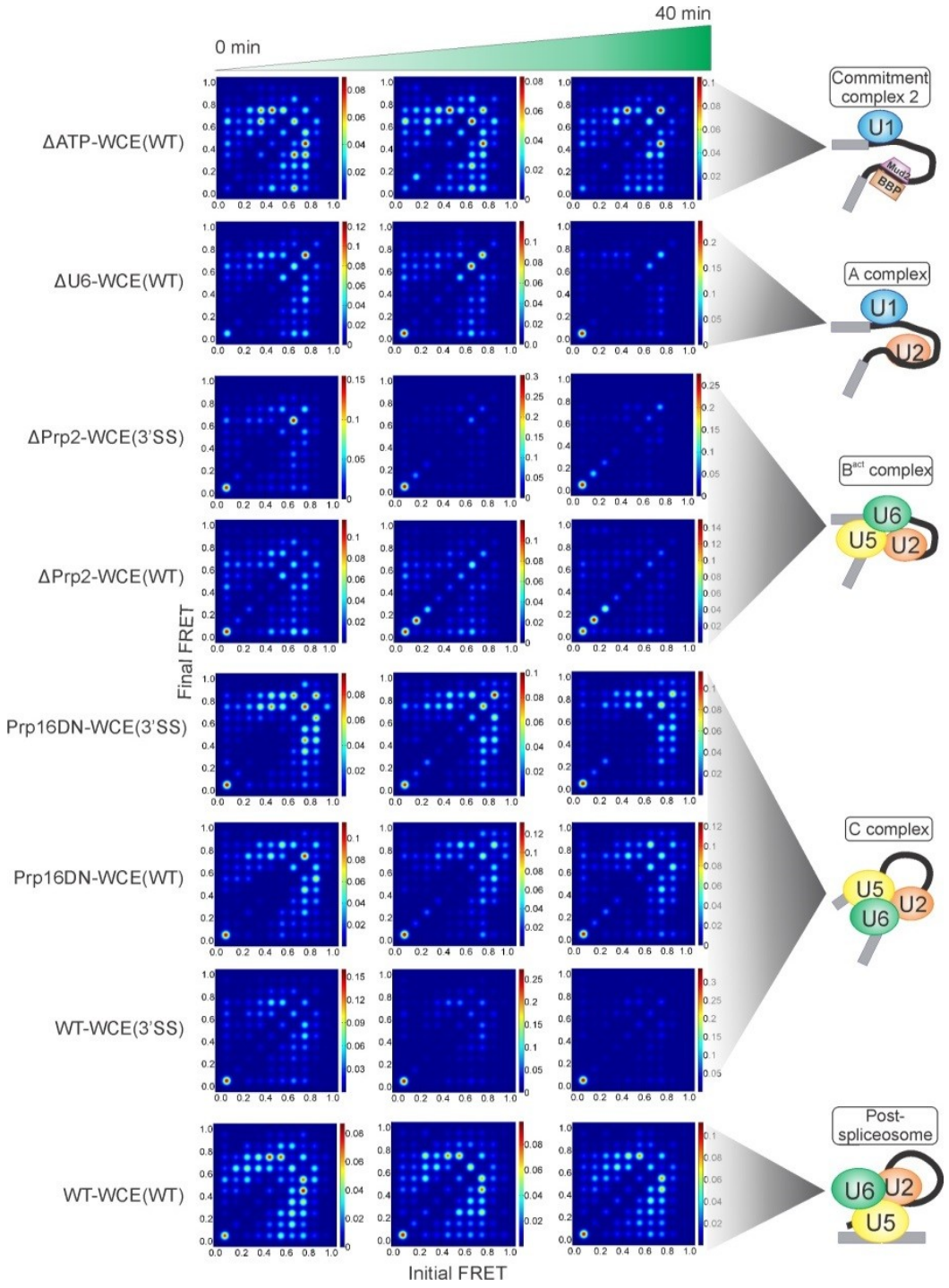



**Figure 3.9 Transition Occupancy Density Plot (TODP) analysis**
TODP analysis for each of the 8 experimental conditions over the time course of the smFRET experiments.

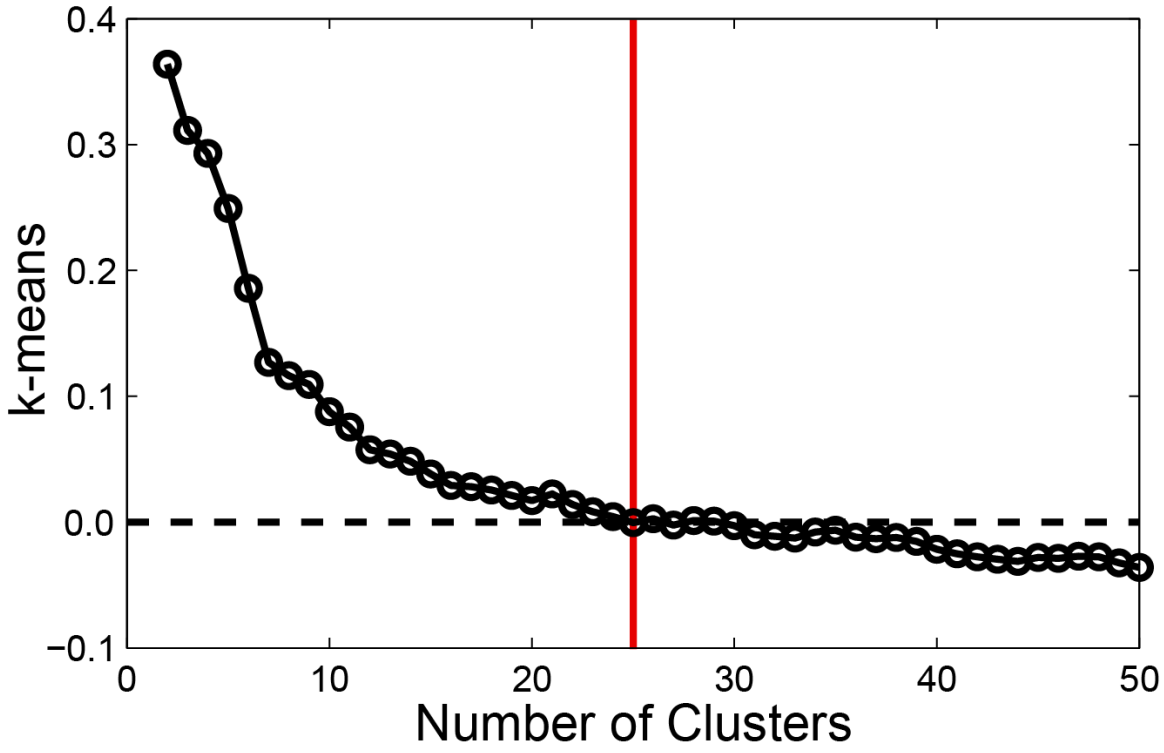
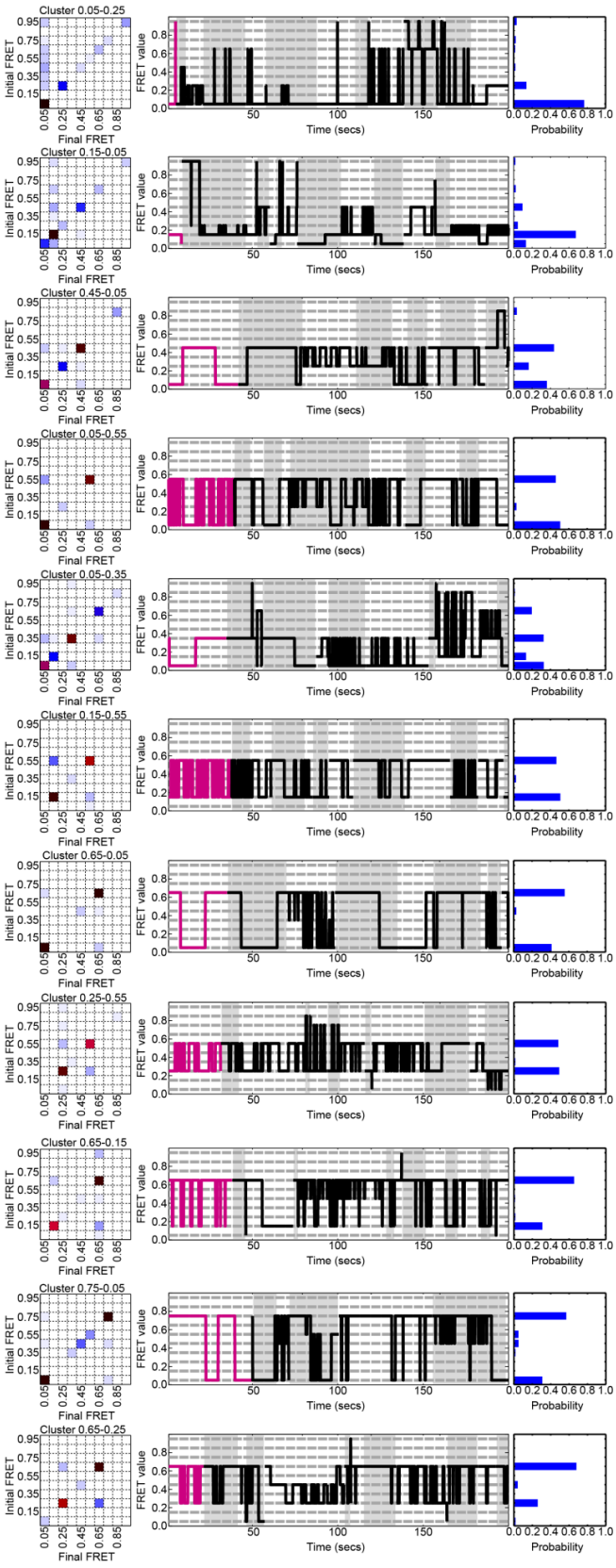
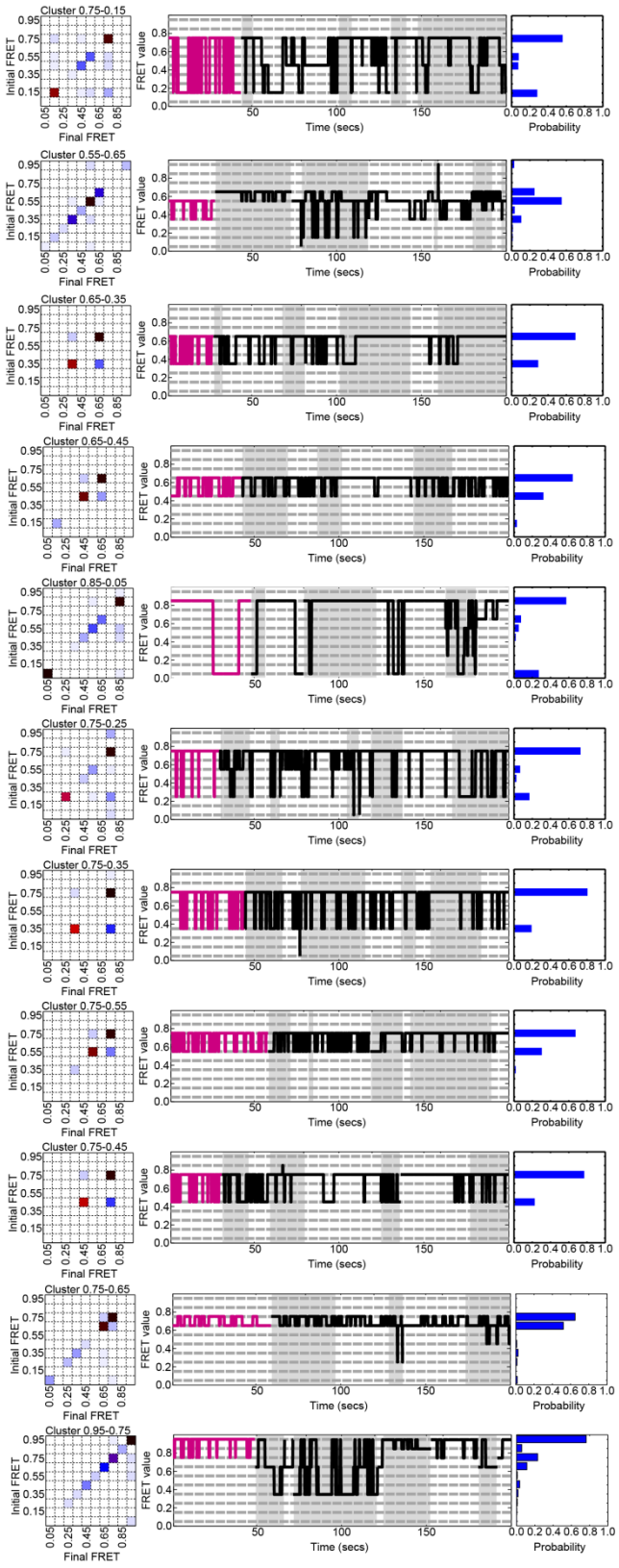



**Figure 3.10 K-means analysis of the optimal cluster number for the full dataset**
Iterative measurement of inter-cluster distances using a modified k-means algorithm utilized to determine the number of clusters that best describes the experimental data.


Cluster 0.05-0.25
Cluster 0.15-0.05
Cluster 0.45-0.05
Cluster 0.05-0.55
Cluster 0.05-0.35
Cluster 0.15-0.55
Cluster 0.65-0.05
Cluster 0.25-0.55
Cluster 0.65-0.15
Cluster 0.75-0.05
Cluster 0.65-0.25
Initial FRET
Final FRET
FRET value
Time (secs)
Probability


Cluster 0.75-0.15
Cluster 0.55-0.65
Cluster 0.65-0.35
Cluster 0.65-0.45
Cluster 0.85-0.05
Cluster 0.75-0.25
Cluster 0.75-0.35
Cluster 0.75-0.55
Cluster 0.75-0.45
Cluster 0.75-0.65
Cluster 0.95-0.75
Initial FRET
Final FRET
FRET value
Time (secs)
Probability

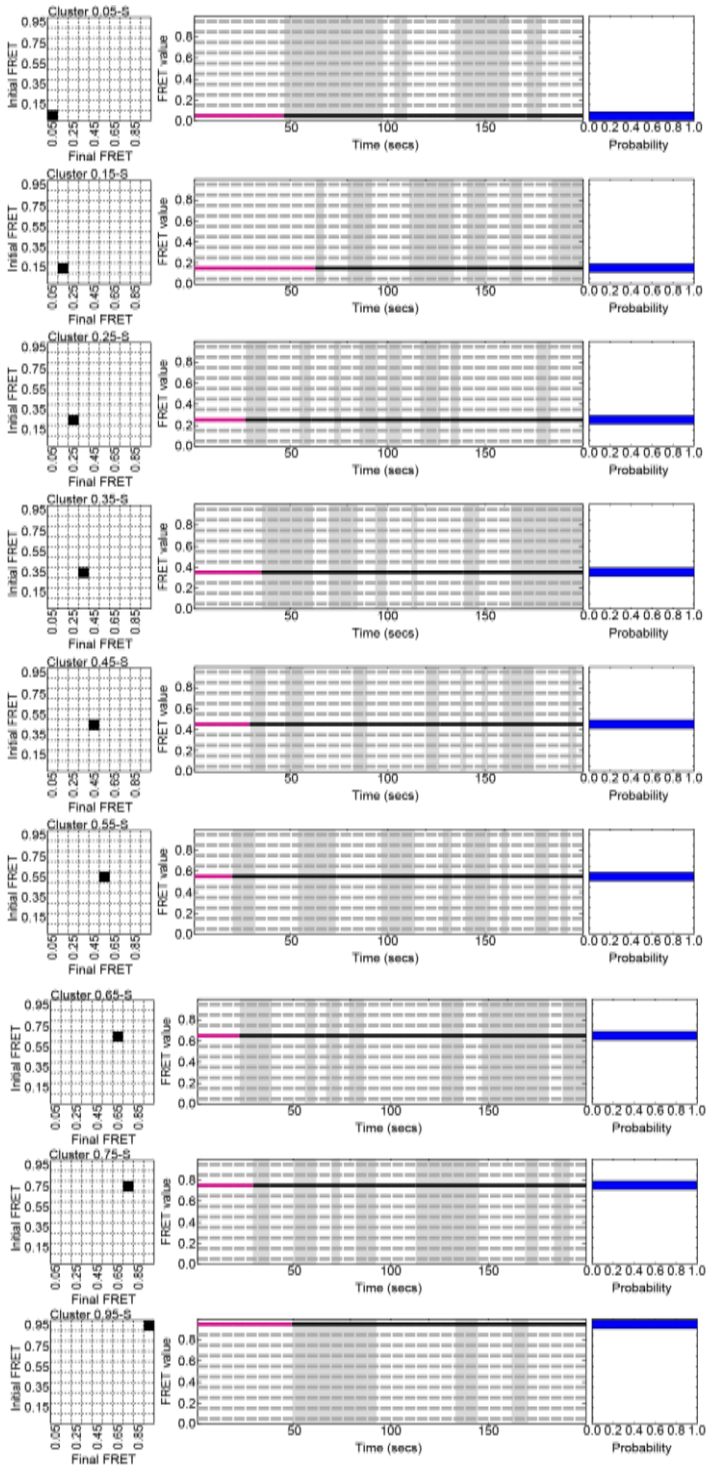



**Figure 3.11 Cluster descriptions**
Transition probability matrix (left), the longest of the 10 traces whose HMM is most similar to the average HMM of the cluster (magenta) and 200 s of random traces (black, middle), and the probability of FRET states for each dynamic and static cluster (right). Grey and white backgrounds demarcate individual trajectories.

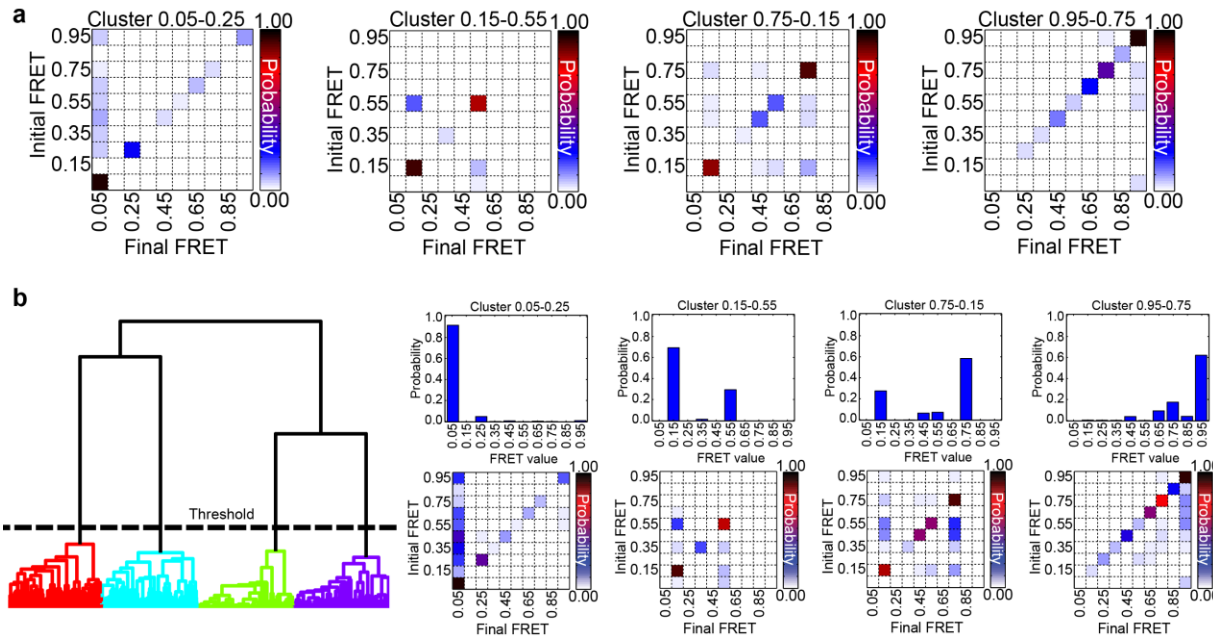



**Figure 3.12 Clustering of simulated dataset produced from four of the dynamic clusters representing the large experimental dataset**

**(a)** The four transition probability (TP) matrices from clustering of the full splicing dataset that were used to generate 1500 random traces for each cluster (see online methods). These traces were pooled and used as input for clustering by SiMCAn. **(b)** Hierarchical tree showing the four cluster threshold found by SiMCAn and the four resulting cluster probability histograms and TP matrices.

SiMCAn-identified clusters for the large experimental dataset capture the molecular behavior exhaustively.

Next, common FRET states and salient kinetic features at each step of the splicing reaction were identified by evaluating the fraction of molecules belonging to each cluster (i.e., the relative cluster occupancy) for each of the eight experimental conditions. We then sought to identify clusters whose occupancies are similarly either enriched or depleted for the same group of conditions, i.e., follow a similar pattern of high and low occupancies across conditions, suggesting they can be grouped into a 'clade' through a second round of hierarchical clustering (**Figure 3.13b**). Upon application of this second level of SiMCAn to the full dataset, a tree height of seven clades (**Figure 3.14**) allowed for the identification of clusters representative of particular splicing conditions, thus most naturally capturing the changes in dynamic behavior expected to occur as the pre-mRNA progresses through the splicing cycle (**Figure 3.15** and **Figure 3.13c**). The fraction of molecules within each cluster for each of our three experimental time points (early, middle, and late) was normalized to a mean occupancy of zero to render differences and similarities among cluster occupancies for each splicing block most evident. A bar graph of all 35 (25 dynamic plus 10 static) clusters was also constructed, revealing the extent to which each cluster contributes to the overall dynamics for each condition (**Figure 3.16** and **Figure 3.17** and **Figure 3.18**). Statistical analysis found that the average length of molecules within each cluster was similar, indicating that SiMCAn does not segregate by trace length (**Figure 3.19** and **Table 3.3**).

### 3.3.6 Characterization of pre- and post-first step splicing blocks

Application of SiMCAn revealed a disperse set of dynamics and cluster occupancies in the early splicing conditions ΔATP-WCE(WT) and ΔU6-WCE(WT) that stall at the Commitment

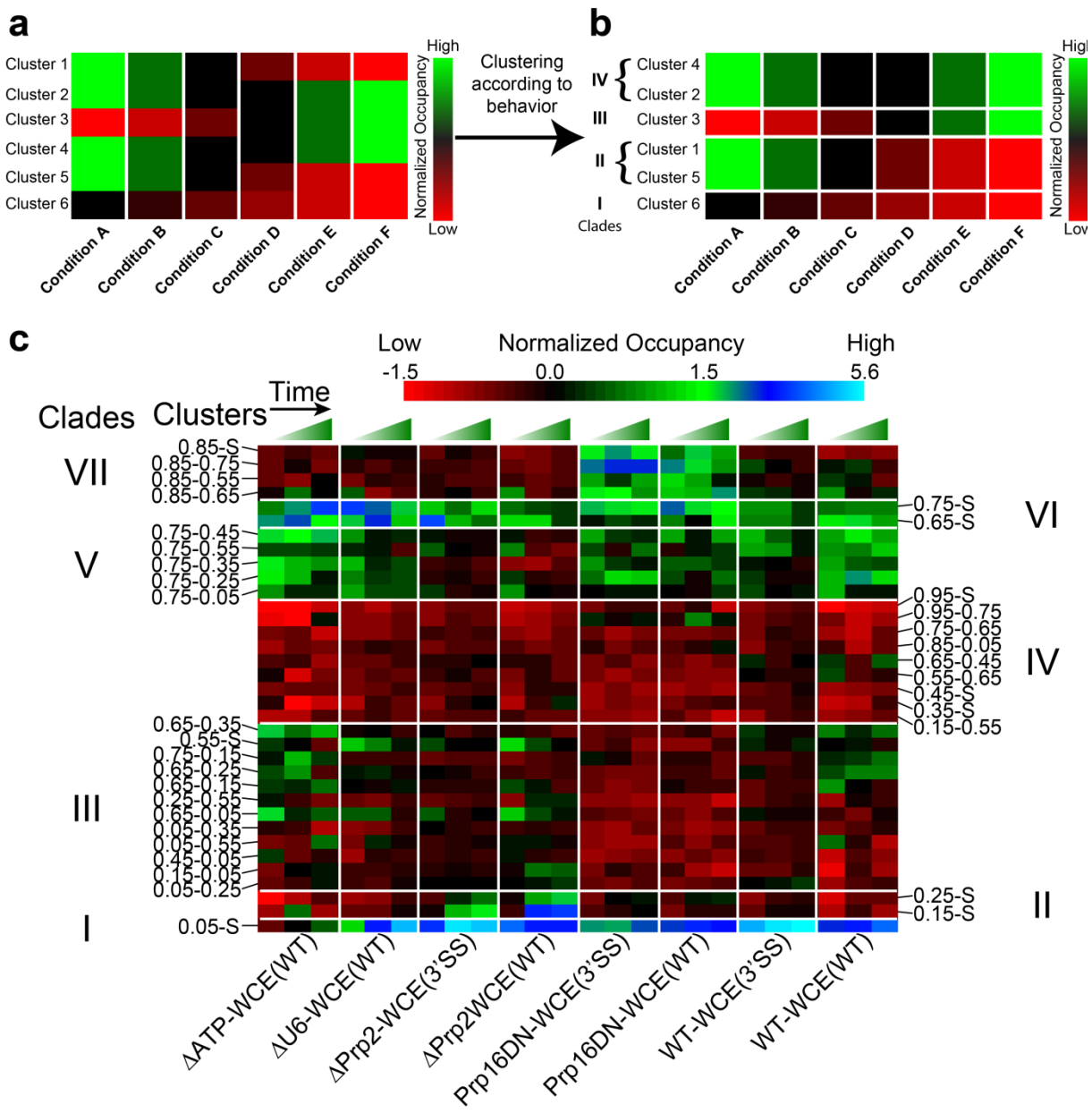



**Figure 3.13 Clustering of clusters to identify 'clades' of similar behavior**
(**a**) Example clustering of smFRET trajectories from three experimental conditions into six clusters. A heat-map representation shows clusters with a large number of molecules (high occupancy) in green and small number of molecules (low occupancy) in red. (**b**) Example clustering of clusters to identify clusters that partition similarly among the three experimental conditions. Clusters that show similar occupancy among the conditions are grouped to form a clade. **(c)** Heat-map representation of the clustering of clusters for the 8 experimental conditions.

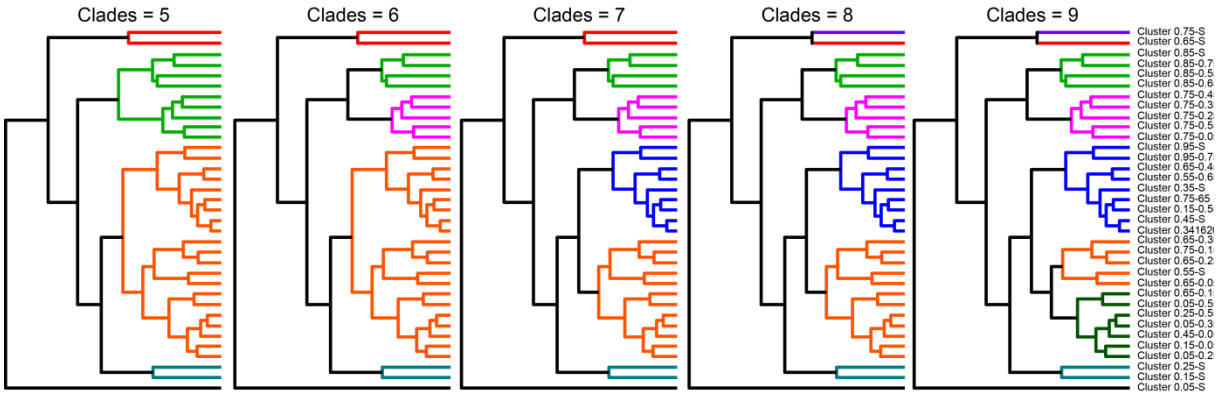



**Figure 3.14 Clade cut-off determination**
Varying the tree cut-off heights upon grouping the cluster occupancy among the 8 experimental conditions leads to distinct numbers of (color-coded) clades of clusters (as indicated on the right).

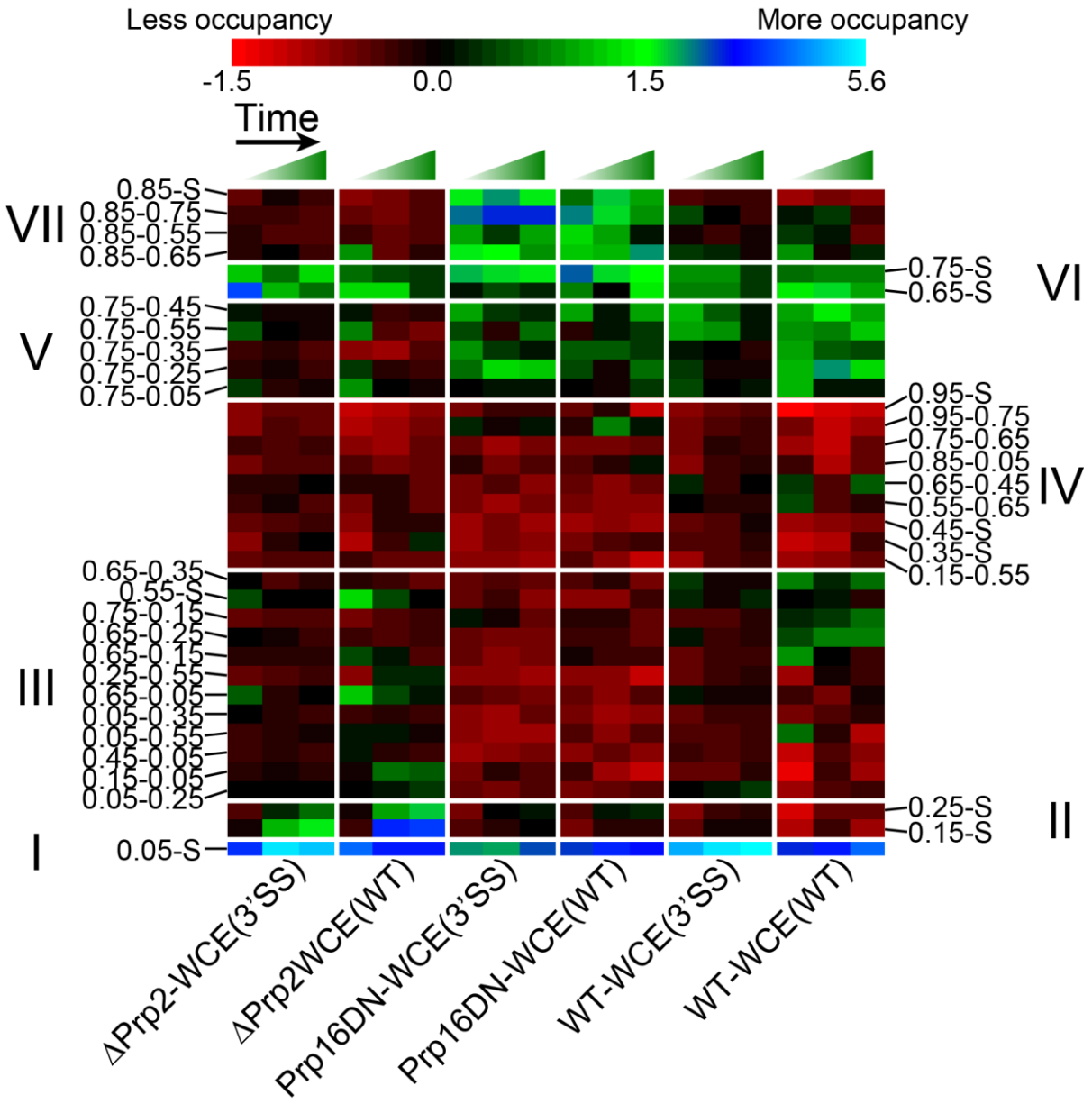



**Figure 3.15 Clades of clusters are enriched in each splicing reaction condition**
Grouping clusters based on their occupancy among the experimental conditions reveals 7 clades (labeled I-VII) of clusters enriched in particular splicing complexes. The fraction of molecules within each cluster for each experimental condition at each time was normalized to a mean of zero with unit variance. Green and blue colors indicate increased occupancy of a particular cluster while red indicates decreased occupancy. Rows identify the clusters and are ordered by increasing average FRET of the clade. Columns identify the cluster occupancy of each condition for the early, middle, and late time points.

Complex 2 (CC2) and A complexes, respectively (**Figure 3.18**). Interestingly, SiMCAn identified a time-dependent increase in clade I upon A complex formation (**Supplementary Note 1**). This low-FRET behavior has been proposed to be sustained during incorporation of the U5·U4/U6 tri-snRNP into the B complex[54], which further progresses through the removal of the U1 and U4 snRNPs upon formation of the activated spliceosome $B^{act}$, a complex enriched through depletion of Prp2[38,68] (**Figure 3.1a)**. In our corresponding ΔPrp2-WCE(WT) and ΔPrp2-WCE(3SS) datasets, SiMCAn recognized a pair of static clusters, 0.25-**S** and 0.15-**S,** found to be overrepresented and thus grouped to form clade II (**Figure 3.15** and **Figure 3.16**). These clusters represent molecules that are stalled in a static low-FRET $B^{act}$ conformation prior to activation of Prp2's ATPase activity (**Figure 3.5**). These FRET states are of slightly different value than those previously determined[55] using the isolated $B^{act}$ complex lacking free extract (0.2 versus 0.3). This is possibly the result of the lower signal-to-noise ratio often associated with the presence of a high concentration of WCE. Alternatively, the proteins and RNAs within the WCE may be inducing a more open conformation of the spliceosome. Notably, SiMCAn was able to distinguish these clusters from the equally static, but even lower FRET cluster 0.05-**S** of the A complex, which is not resolvable in the FRET histograms (**Figure 3.8**). In addition to the static clusters of clade II, the dynamic cluster 0.05-0.25 (**Figure 3.20a**) is moderately enriched in these conditions relative to other conditions, suggesting that occasional excursions back into an A or B-like conformation occur.

In contrast to Prp2 depletion, SiMCAn identified clade VII as particularly enriched upon addition of recombinant Prp16 dominant negative mutant ATPase (Prp16DN-WCE(WT) and Prp16DN-WCE(WT)), known to stall splicing within the post-first-step C complex[55,127,128] (**Figure 3.15, Figure 3.16, and Figure 3.20).** Within this clade were a static cluster 0.85-S and

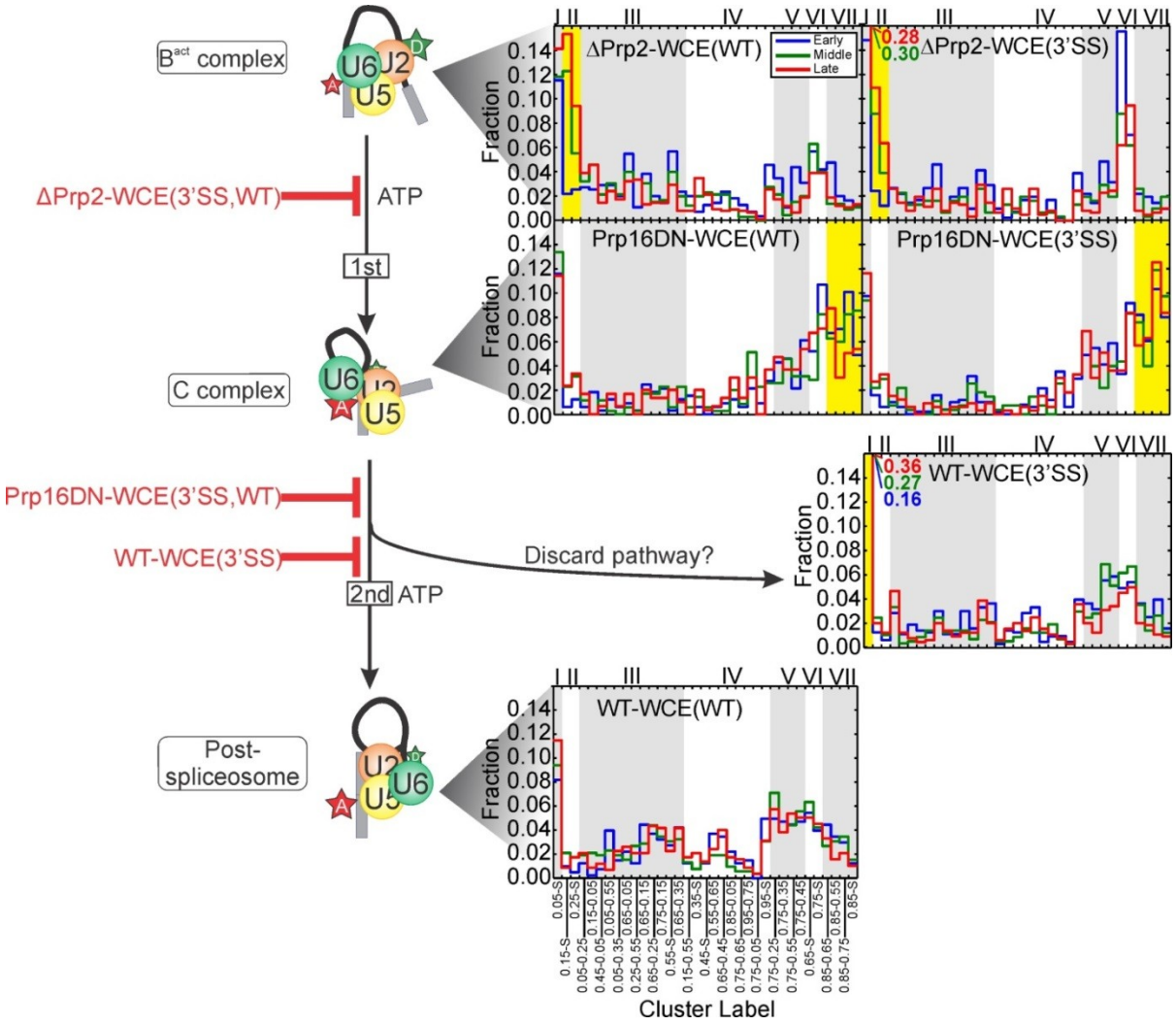



**Figure 3.16 Cluster occupancy histogram post-first step splicing blocks**
Cluster occupancy histogram showing the raw fraction of molecules occupying each cluster for the late assembly stages of the splicing cycle. Alternating gray and white backgrounds demarcate the clusters (bottom) comprising each of the 7 clades (top). Clusters of significant occupancy within a specified condition are highlighted in yellow.

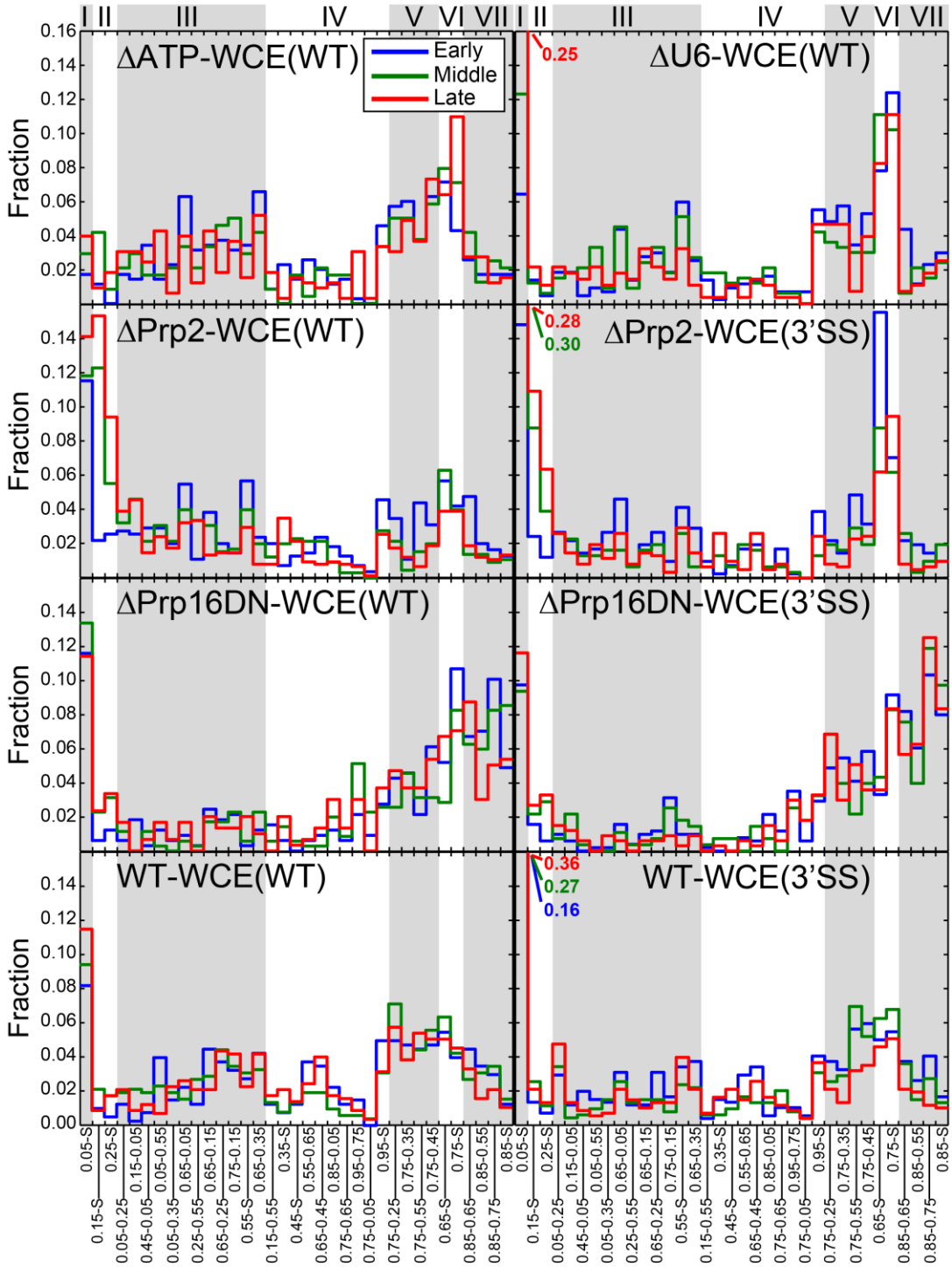



**Figure 3.17 Experimental datasets show vastly different cluster occupancies**
Histogram showing the raw fraction of molecules occupying each cluster of the 8 experimental conditions. Alternating gray and white backgrounds demarcate the clusters (bottom) comprising each of the 7 clades (top).

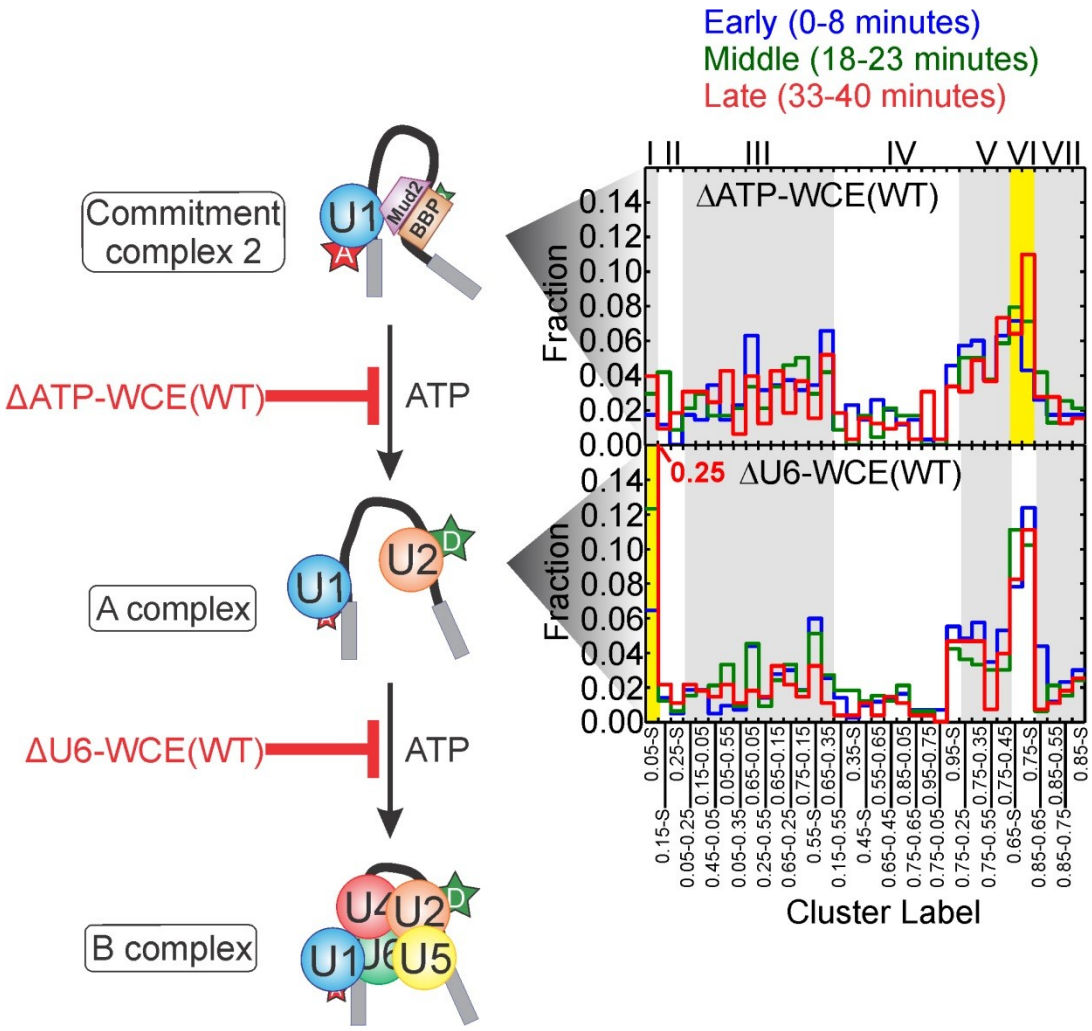
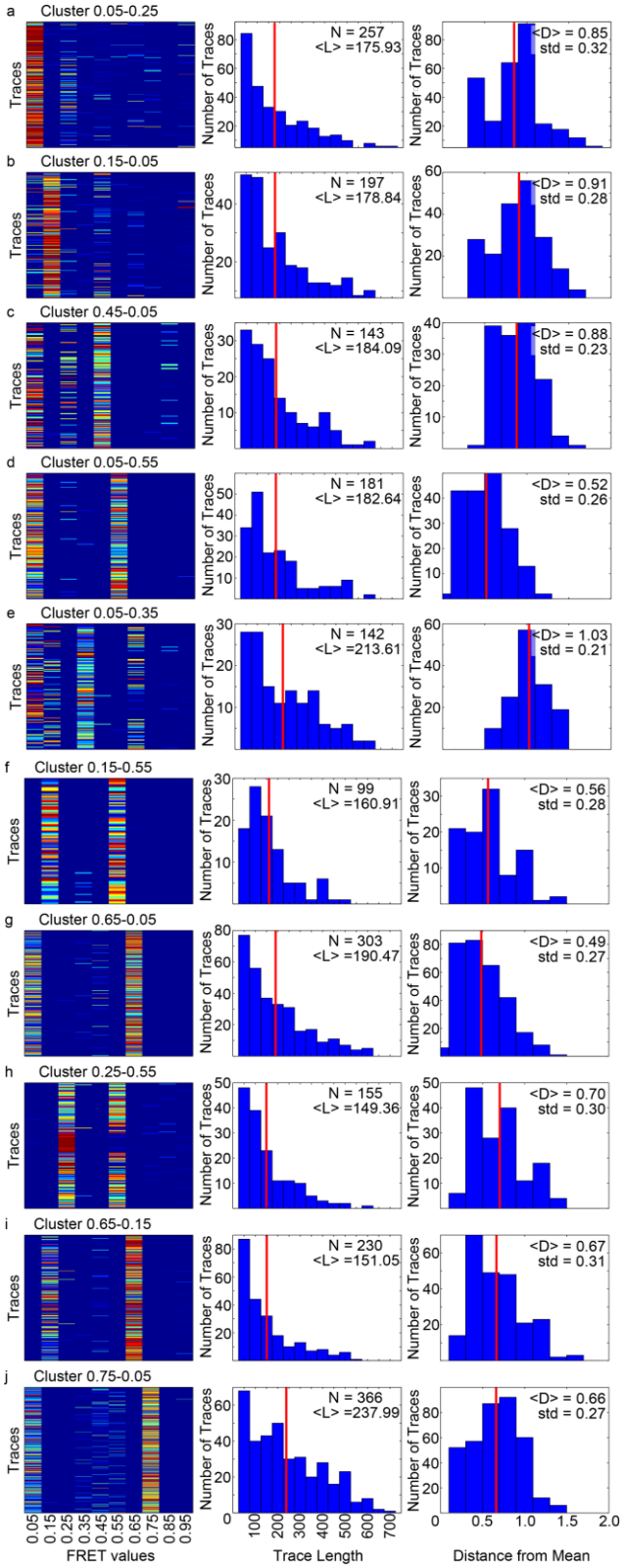


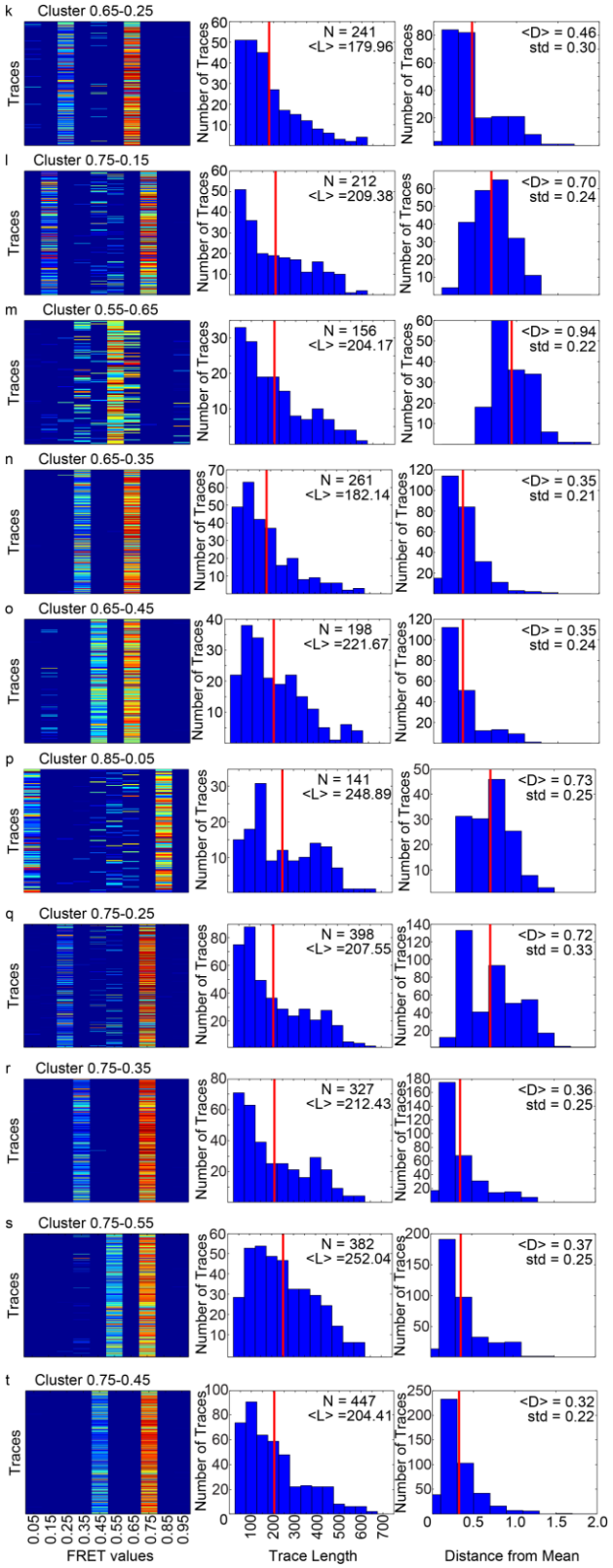

**Figure 3.18 Early splicing blocks show a dramatic shift in cluster occupancy**
Histogram showing the raw fraction of molecules occupying each cluster for the early splicing block conditions. Alternating gray and white backgrounds demarcate the clusters (bottom) comprising each of the 7 clades (top). Clusters of significant occupancy within a specified condition are highlighted in yellow.

three dynamic clusters, all containing the 0.85 FRET state (**Figure 3.21**), which is distinct from the 0.75-S/0.65-S conformational state of clade VI enriched in early splicing intermediates (**Figure 3.13c** and **Figure 3.18**). The dynamics of the clusters enriched at the Prp16DN stage indicate a preference for the 0.85 high-FRET state (**Figure 3.21b**), suggesting we are enriching for and identifying molecules just before catalysis or transiently sampling the first catalytic conformation before proceeding to the 0.85-S cluster characteristic of molecules that have undergone first-step splicing. Although the ΔPrp2-WCE(3'SS) stall did show a delay in $B^{act}$ complex formation (**Supplementary Note 2**), these observations suggest that only faithful spliceosome assembly leads to juxtaposition of the 5'SS and BP in a stable fashion, thus favoring first-step catalysis independent of the identity of the 3'SS.

3.3.7 A 3'SS mutation leads to undocking late in spliceosome assembly

Finally, SiMCAn identified differences in smFRET behavior between the WT and 3'SS mutant substrates upon incubation with WT WCE containing no blocks (WT-WCE(WT) and WT-WCE(3'SS)), thus allowing for the unabated assembly towards the final step of splicing. The 3'SS mutant is known to assemble in a complex that includes the splicing factors responsible for the second step of catalysis, yet the 3'SS mutant is not amenable to splicing (**Figure 3.2**). Since both substrates progress through most of the splicing cycle, it is not surprising that SiMCAn revealed a similar set of pre-mRNA conformations sampled (**Figure 3.16**). However, the 3'SS over time adopted an increasingly dominant 0.05-**S** cluster (**Figure 3.16,** clade I), indicating a large separation of the 5'SS and BP not found in the Prp16DN-WCE(3'SS) dataset. This 0.05-**S** state is thus stabilized to a much greater extent in the 3'SS mutant than the WT substrate, supporting the appearance of a conformation in which the 5'SS and BP become greatly separated only after the first step of splicing when the mutated 3'SS is detected. Our data suggest that the


a
Cluster 0.05-0.25
N = 257
<L> =175.93
<D> = 0.85
std = 0.32
b
Cluster 0.15-0.05
N = 197
<L> =178.84
<D> = 0.91
std = 0.28
c
Cluster 0.45-0.05
N = 143
<L> =184.09
<D> = 0.88
std = 0.23
d
Cluster 0.05-0.55
N = 181
<L> =182.64
<D> = 0.52
std = 0.26
e
Cluster 0.05-0.35
N = 142
<L> =213.61
<D> = 1.03
std = 0.21
f
Cluster 0.15-0.55
N = 99
<L> =160.91
<D> = 0.56
std = 0.28
g
Cluster 0.65-0.05
N = 303
<L> =190.47
<D> = 0.49
std = 0.27
h
Cluster 0.25-0.55
N = 155
<L> =149.36
<D> = 0.70
std = 0.30
i
Cluster 0.65-0.15
N = 230
<L> =151.05
<D> = 0.67
std = 0.31
j
Cluster 0.75-0.05
N = 366
<L> =237.99
<D> = 0.66
std = 0.27
Traces
Number of Traces
FRET values
Trace Length
Distance from Mean


k
Cluster 0.65-0.25
N = 241
<L> =179.96
<D> = 0.46
std = 0.30
l
Cluster 0.75-0.15
N = 212
<L> =209.38
<D> = 0.70
std = 0.24
m
Cluster 0.55-0.65
N = 156
<L> =204.17
<D> = 0.94
std = 0.22
n
Cluster 0.65-0.35
N = 261
<L> =182.14
<D> = 0.35
std = 0.21
o
Cluster 0.65-0.45
N = 198
<L> =221.67
<D> = 0.35
std = 0.24
p
Cluster 0.85-0.05
N = 141
<L> = 248.89
<D> = 0.73
std = 0.25
q
Cluster 0.75-0.25
N = 398
<L> =207.55
<D> = 0.72
std = 0.33
r
Cluster 0.75-0.35
N = 327
<L> =212.43
<D> = 0.36
std = 0.25
s
Cluster 0.75-0.55
N = 382
<L> =252.04
<D> = 0.37
std = 0.25
t
Cluster 0.75-0.45
N = 447
<L> =204.41
<D> = 0.32
std = 0.22
Traces
Number of Traces
FRET values
Trace Length
Distance from Mean

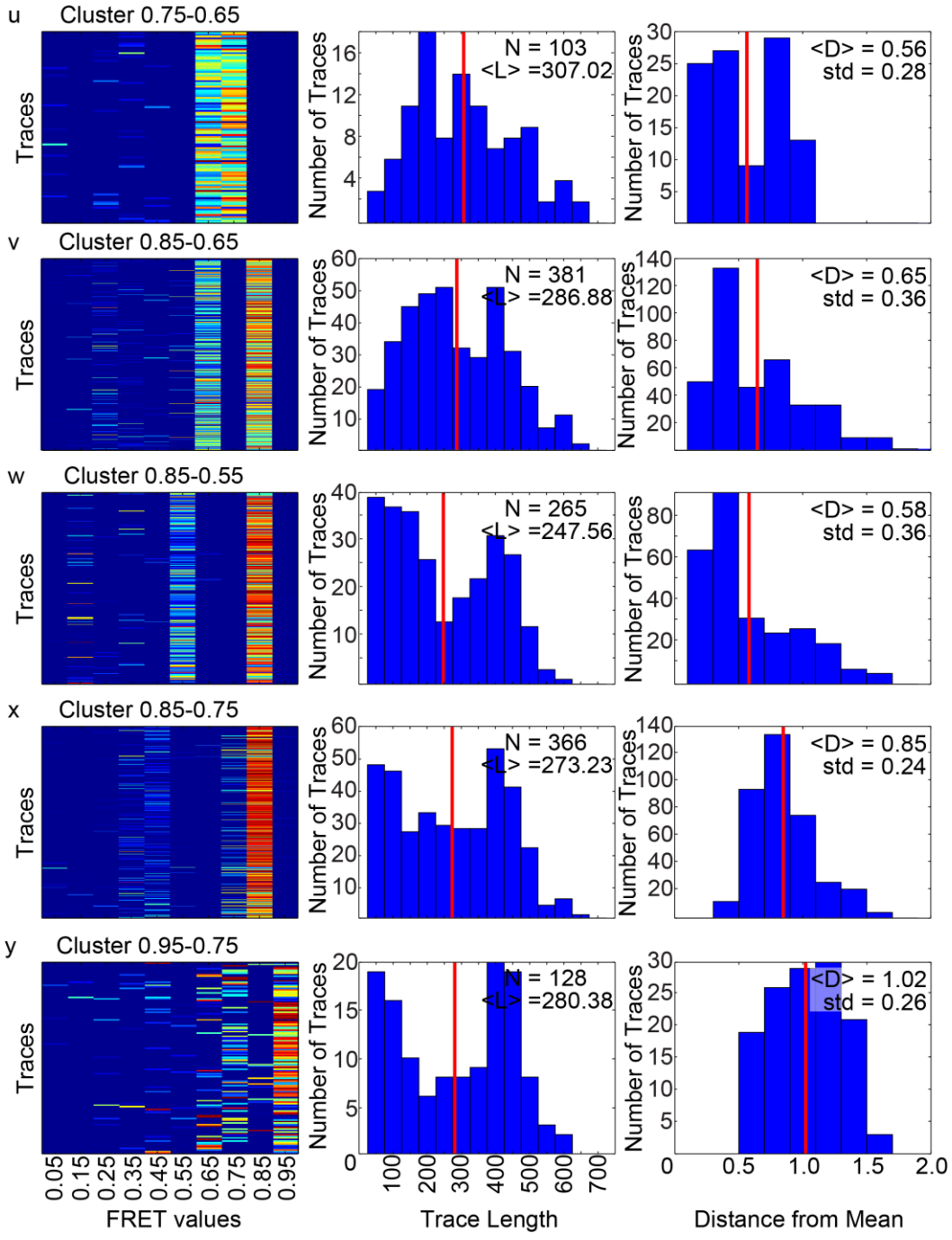



**Figure 3.19 Statistical analysis of all 35 (25 dynamic, 10 static) clusters**
**(a-y)** The left panel depicts every trace that contributes to a cluster with the heat map indicating the occupancy at the state (blue = 0, red = 1). The middle panel indicates the distribution of trace length in each cluster. The number of molecules in each cluster (N) and the average trace length (<L>, red line) are indicated in the top right corner. The right panel plots the distance of each trace's HMM from the mean HMM of the cluster. The average distance (<D>, red line) and standard deviation (std) are indicated.

| Cluster label | mean fret value | number of traces | mean length | std Length | Mean distance from centroid | std distance |
|---|---|---|---|---|---|---|
| 0.05-0.25 | 0.14 | 257 | 175.93 | 139.10 | 0.85 | 0.32 |
| 0.15-0.05 | 0.20 | 197 | 178.84 | 135.54 | 0.91 | 0.28 |
| 0.45-0.05 | 0.29 | 143 | 184.09 | 130.64 | 0.88 | 0.23 |
| 0.05-0.55 | 0.29 | 181 | 182.64 | 134.64 | 0.52 | 0.26 |
| 0.05-0.35 | 0.29 | 142 | 213.61 | 145.09 | 1.03 | 0.21 |
| 0.15-0.55 | 0.34 | 99 | 160.91 | 105.35 | 0.57 | 0.28 |
| 0.65-0.05 | 0.40 | 303 | 190.47 | 140.47 | 0.49 | 0.27 |
| 0.25-0.55 | 0.40 | 155 | 149.36 | 113.54 | 0.70 | 0.30 |
| 0.65-0.15 | 0.49 | 230 | 151.05 | 122.23 | 0.67 | 0.31 |
| 0.75-0.05 | 0.50 | 366 | 237.99 | 157.11 | 0.66 | 0.27 |
| 0.65-0.25 | 0.53 | 241 | 179.96 | 128.72 | 0.46 | 0.30 |
| 0.75-0.15 | 0.54 | 212 | 209.38 | 148.37 | 0.70 | 0.24 |
| 0.55-0.65 | 0.55 | 156 | 204.17 | 144.34 | 0.94 | 0.22 |
| 0.65-0.35 | 0.56 | 261 | 182.14 | 125.20 | 0.35 | 0.21 |
| 0.65-0.45 | 0.57 | 198 | 221.67 | 135.07 | 0.35 | 0.24 |
| 0.85-0.05 | 0.59 | 141 | 248.89 | 147.61 | 0.73 | 0.25 |
| 0.75-0.25 | 0.64 | 398 | 207.55 | 146.12 | 0.72 | 0.33 |
| 0.75-0.35 | 0.67 | 327 | 212.43 | 148.21 | 0.36 | 0.25 |
| 0.75-0.55 | 0.68 | 382 | 252.04 | 140.19 | 0.37 | 0.25 |
| 0.75-0.45 | 0.68 | 447 | 204.41 | 137.48 | 0.32 | 0.22 |
| 0.75-0.65 | 0.68 | 103 | 307.02 | 144.65 | 0.56 | 0.28 |
| 0.85-0.65 | 0.75 | 381 | 286.88 | 141.36 | 0.65 | 0.36 |
| 0.85-0.55 | 0.75 | 265 | 247.56 | 149.19 | 0.58 | 0.36 |
| 0.85-0.75 | 0.79 | 366 | 273.23 | 154.33 | 0.85 | 0.24 |
| 0.95-0.75 | 0.85 | 128 | 280.38 | 162.00 | 1.02 | 0.26 |
| 0.05-S | 0.05 | 1601 | 172.35 | 124.06 | 0.00 | 0.00 |
| 0.15-S | 0.15 | 438 | 142.36 | 95.30 | 0.00 | 0.00 |
| 0.25-S | 0.25 | 277 | 123.81 | 87.50 | 0.00 | 0.00 |
| 0.35-S | 0.35 | 138 | 115.83 | 85.62 | 0.00 | 0.00 |
| 0.45-S | 0.45 | 125 | 93.06 | 67.01 | 0.00 | 0.00 |
| 0.55-S | 0.55 | 317 | 104.01 | 71.82 | 0.00 | 0.00 |
| 0.65-S | 0.65 | 656 | 119.75 | 86.34 | 0.00 | 0.00 |
| 0.75-S | 0.75 | 722 | 146.99 | 107.65 | 0.00 | 0.00 |
| 0.85-S | 0.85 | 284 | 180.22 | 128.57 | 0.00 | 0.00 |
| 0.95-S | 0.95 | 43 | 276.91 | 136.23 | 0.00 | 0.00 |

**Table 3.3 Statistical analysis of each of the 35 clusters**

3'SS is either unable to dock into the catalytic core or is unable to remain docked in the catalytic core after the ATP-dependent action of Prp16. This deficiency in docking may be a result of second-step factors preventing docking into the second-step conformation[129,130]. Alternatively, this open conformation may be caused by Prp22, an ATPase known to be involved in proofreading mutant substrates during the second step of splicing (**Supplementary Note 3**)[43,116]. In this latter case, the 3'SS may transiently dock into the second-step conformation, but Prp22 rapidly recognizes and discards the mutated 3'SS. Either hypothesis would explain the accumulation of a discarded, undocked substrate unable to proceed through the second step of splicing. Taken together, our SiMCAn analysis suggests that the lack of a proper 3'SS sequence marker leads to robust proofreading against a substrate incompetent for the second step of splicing by undocking from the active site.

### 3.4 Discussion

We here have demonstrated the power of Single Molecule Cluster Analysis (SiMCAn) as applied to a large smFRET dataset collected on the spliceosome in the presence of various biochemical and genetic stalls. By coupling SiMCAn with multiple experimental conditions of defined impact on the splicing cycle, we show that such an smFRET dataset can be efficiently clustered and analyzed to reveal unique dynamic properties associated with specific splicing cycle intermediates that could not be identified using classical histogram and TODP analysis (**Figure 3.8** and **Figure 3.9**). Since SiMCAn does not make assumptions about the heterogeneity or completeness of the underlying biochemical reactions, it allows one to identify consistent molecular behaviors in model-free fashion. Through such unbiased and thorough analysis we were able to assign dynamic FRET states to specific complexes, identify molecules transitioning between complexes, and demonstrate that the 5'SS and BP undock completely after the first step

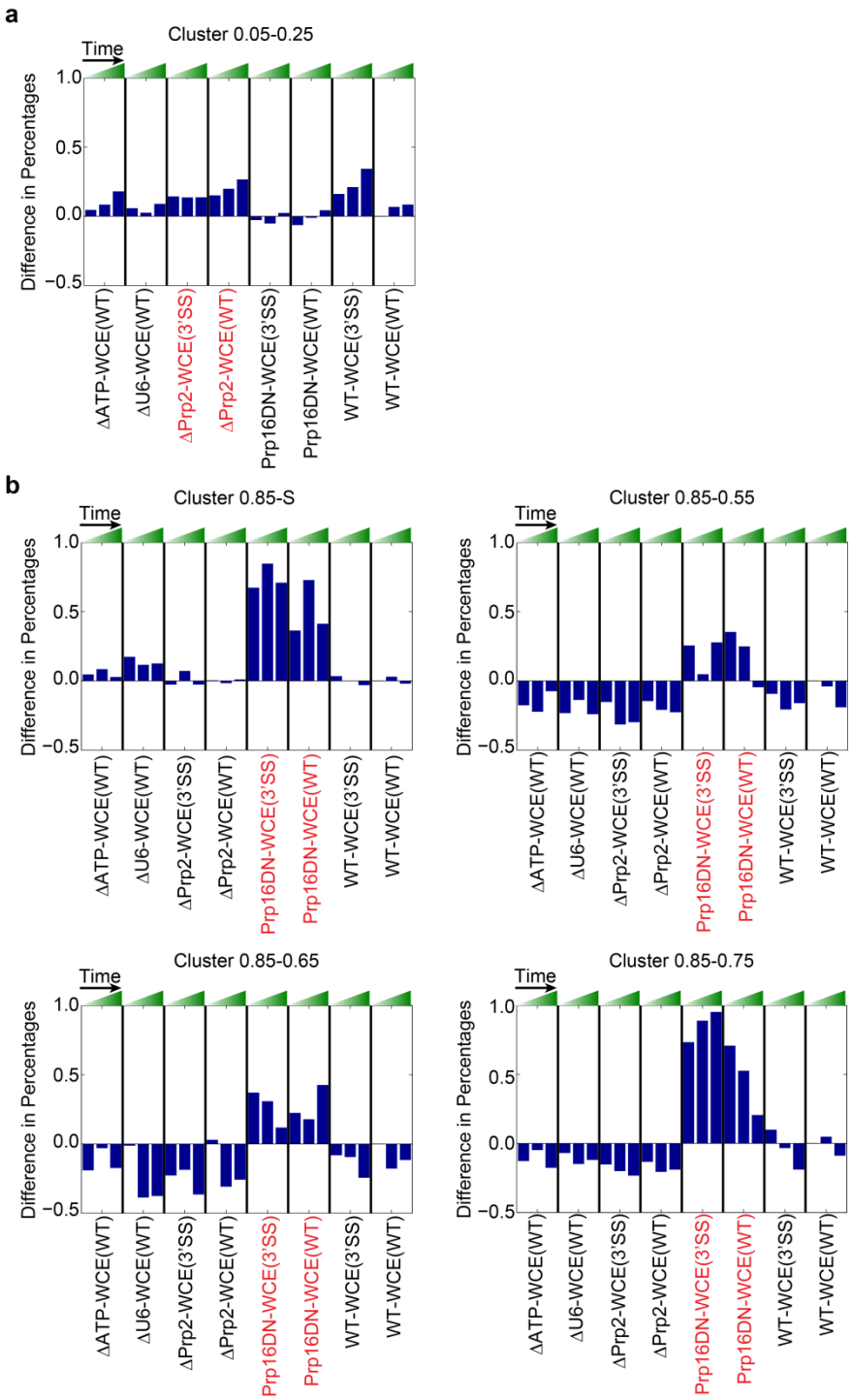



**Figure 3.20 Clusters enriched in particular splicing conditions**
The occupancy of clusters within each of the 8 experimental conditions compared to that of WT-WCE(WT). Each occupancy value for every condition is subtracted from the occupancy of the cluster in the WT-WCE(WT) early condition. **(a)** Cluster enriched in the ΔPrp2-WCE condition (red). **(b)** Clusters enriched in the Prp16DN-WCE condition (red).

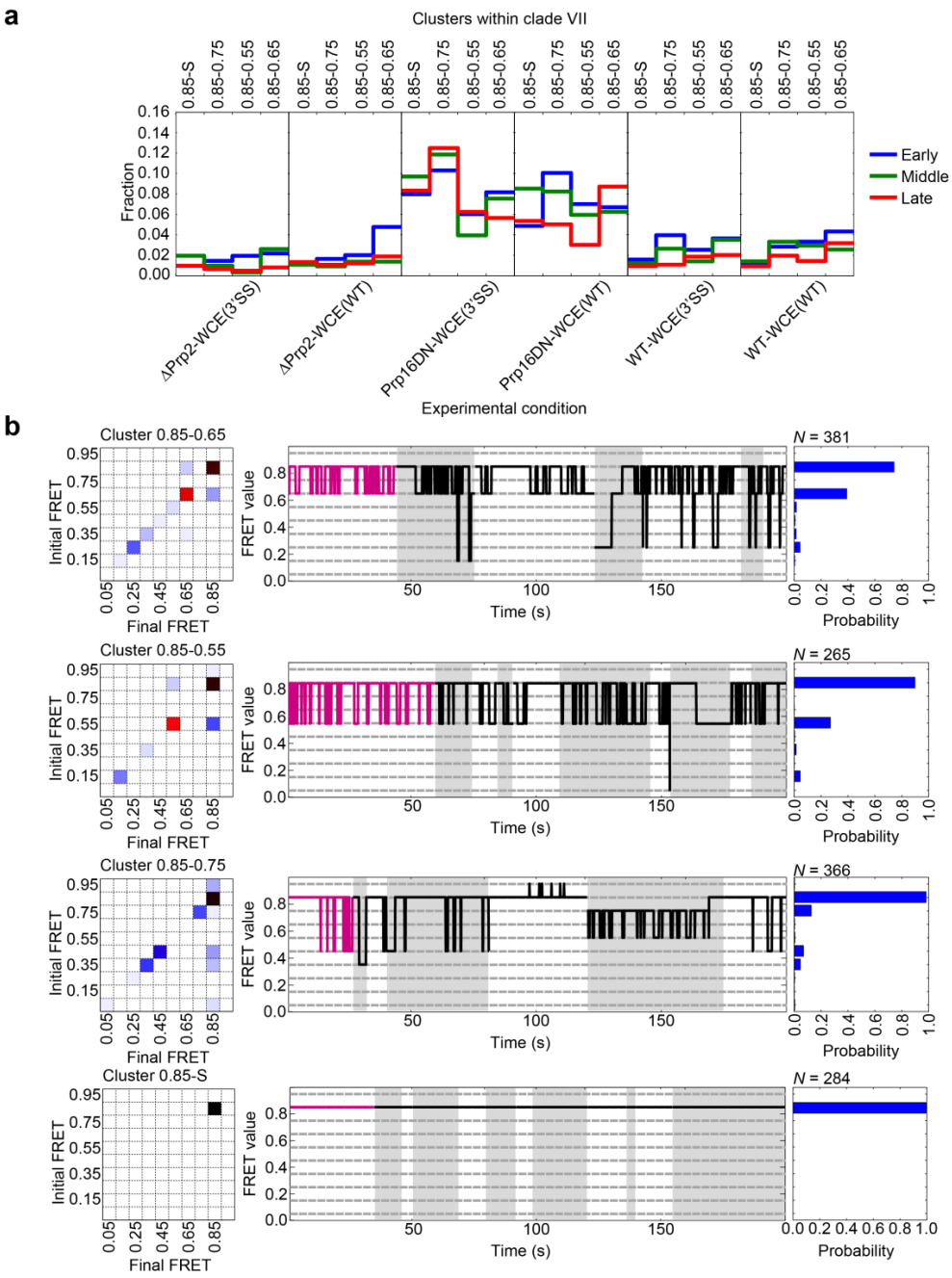



**Figure 3.21 Dynamic clusters of clade VII enriched in the Prp16DN-WCE conditions show repeated excursions from the 0.85 state to lower FRET states.**
**(a)** Fraction of molecules within each late assembly stage for the clusters of clade VII. **(b)** Cluster description for each of the four clusters within clade VII. Each representation shows the TP matrix of the cluster (left), the trace closest to the cluster center (magenta) and up to 200 s of random (black) traces from the cluster (middle), and the probability of FRET states within the cluster (right). Grey and white backgrounds demarcate individual trajectories.

of splicing when the spliceosome encounters a 3'SS mutation (**Figure 3.16**). SiMCAn thus can use exploratory datasets collected from complex reaction pathways to generate testable hypotheses, for example, that the spliceosome exploits similar undocked intermediates to proofread substrates along the splicing cycle, providing checkpoints that trap suboptimal substrates not meeting the criteria for cycle progression.

Single molecule FRET experiments provide a unique perspective into the dynamic behavior of complex reactions like splicing. Our experiments revealed a complex set of dynamic behaviors throughout the splicing cycle. SiMCAn was born of the necessity to classify common kinetic behaviors over a broad range of experimental states. Building hierarchical trees from disparate sets of data is the basis of most phylogenetic inference, and the methods presented here are inspired from evolutionary analysis[131]. The clades identified by SiMCAn allow us to define common subsets of relative dynamic behavior occurring at different biochemical blocks of the splicing cycle. Building on the phylogenetic analogy, the dynamic clades identified represent common kinetic pathways traversing the splicing cycle. We thus observed conserved pathways in the splicing cycle driven by a limited number of transitions. A limitation of our approach is that it does not allow us to unambiguously define conformations from FRET states. In a simpler system, like the P4-P6 subdomain of the *T. thermophila* group I intron, docking/undocking of the GNRA tetraloop could be assigned to specific FRET values, which enabled an unambiguous kinetic model to be developed[119]. Emerging approaches involving multiple probes such as the coincidence analysis of colocalization single-molecule spectroscopy (CoSMoS)[110], combined with SiMCAn, are poised to resolve this ambiguity and facilitate the development of a complete kinetic model of the eukaryotic splicing cycle. As single molecule techniques are applied to increasingly complex biochemical processes, SiMCAn is an approach that will make it possible

to no longer limit the experimental strategy to one with a low number of states while still seeing the forest for the trees.

In summary, our results demonstrate that SiMCAn vastly improves the amount of information possible to extract from a large quantity of complex smFRET data. It is a powerful tool for the unbiased extraction of FRET states and kinetics from single molecule trajectories. By combining Hidden Markov Models with hierarchical clustering, we have utilized the strengths of both techniques to allow for the identification of biologically related dynamics. Beyond the identification of FRET states, SiMCAn helps distinguish molecules with similar FRET levels but differing rates of interconversion. By applying an additional layer of clustering based on the occupancy of behaviors across a systematic set of experimental conditions with known effects, we have created a tool for the identification of common and distinct behaviors among large numbers of single molecules. As such, SiMCAn can help generate hypotheses that drive focused experiments on isolated pathway intermediates. We anticipate that SiMCAn will be a powerful analysis tool that can be applied to any single molecule dataset, allowing for unprecedented in-depth analyses of the dynamics of complex biomolecular machines.

### 3.5 Supplementary Note 1

We subjected the entire dataset of 8 experimental conditions to global analysis by SiMCAn. Application of SiMCAn revealed a disperse set of dynamics and cluster occupancies in the early splicing conditions ΔATP-WCE(WT) and ΔU6-WCE(WT) that stall at the CC2 and A complexes, respectively (**Figure 3.13c** and **Figure 3.18**). Starting with a condition that favors formation of commitment complex 2 (CC2) by depletion of ATP (ΔATP-WCE(WT)), SiMCAn revealed clusters 0.75-**S** and 0.65-**S** of clade VI as the dominant clusters representing a conformational state that increases over our time course (**Figure 3.13c**), indicating that the 5'SS

and BP of the substrate are in close proximity. Such a behavior is expected for Ubc4 pre-mRNA, which contains a highly secondary structured intron with proximal 5'SS and BP[53]. Given that Ubc4 is able to efficiently form CC2 upon incubation with extract depleted of ATP (**Figure 3.23**), this also suggests that binding of U1 snRNP and BBP/Mud2 in CC2 is not sufficient to disrupt this secondary structure, which places the 5'SS and BP potentially close enough, but not properly positioned, for first-step catalysis. A group of dynamic clusters containing 0.75 and 0.65 as the most dominant states (clades III and V) was also significantly enriched (**Figure 3.18**), potentially signifying reversible binding and unbinding of the U1 snRNP and BBP/Mud2 to the pre-mRNA. However, such binding remains transient without the availability of ATP to activate the DExD/H-box ATPase Prp5 and load the U2 snRNA-protein complex (snRNP) onto the BP.

Accordingly, upon addition of extract containing ATP but depleted of U6 snRNA (ΔU6-WCE(WT)) to favor the A complex, SiMCAn identified a time-dependent increase a low-FRET 0.05-**S** cluster (clade I), indicating disruption of Ubc4's secondary structure and undocking of its 5'SS and BP (**Figure 3.13c and Figure 3.18)**. This finding is consistent with the proposal that pre-mRNAs do not sample a proximal 5'SS-BP conformation until a later stage in spliceosome assembly after incorporation of the U5·U4/U6 tri-snRNP upon formation of the activated spliceosome ($B^{act}$ complex)[55]. Notably, the preceding CC2 complex shows low occupancy in the 0.05-**S** cluster (**Figure 3.18**), further supporting the notion that its adoption requires an ATP-dependent assembly event. In this low-FRET state, the 5'SS and BP are stably undocked from one another, preventing premature catalysis prior to proper recognition and proofreading by the spliceosome. The dynamic clusters of clades III and V were again found to be moderately populated (**Figure 3.13c**). Interestingly, several clusters in clade III appear to decrease over time as a result of the increase in occupancy of the 0.05-**S** cluster. This most likely indicates that

reversible excursions are intrinsic to the complex, but can be biased towards a particular conformation upon activation of an ATPase. Furthermore, SiMCAn identified a significant population of molecules under A complex conditions that remained in the 0.75-**S** and 0.65-**S** clusters of clade II, characteristic of CC2 (**Figure 3.18**). It is likely that these molecules were not properly assembled into A complex and remain in a CC2-like state, consistent with the expected incomplete progression through the splicing cycle (**Figure 3.2** and **Figure 3.23**).

### 3.6 Supplementary Note 2

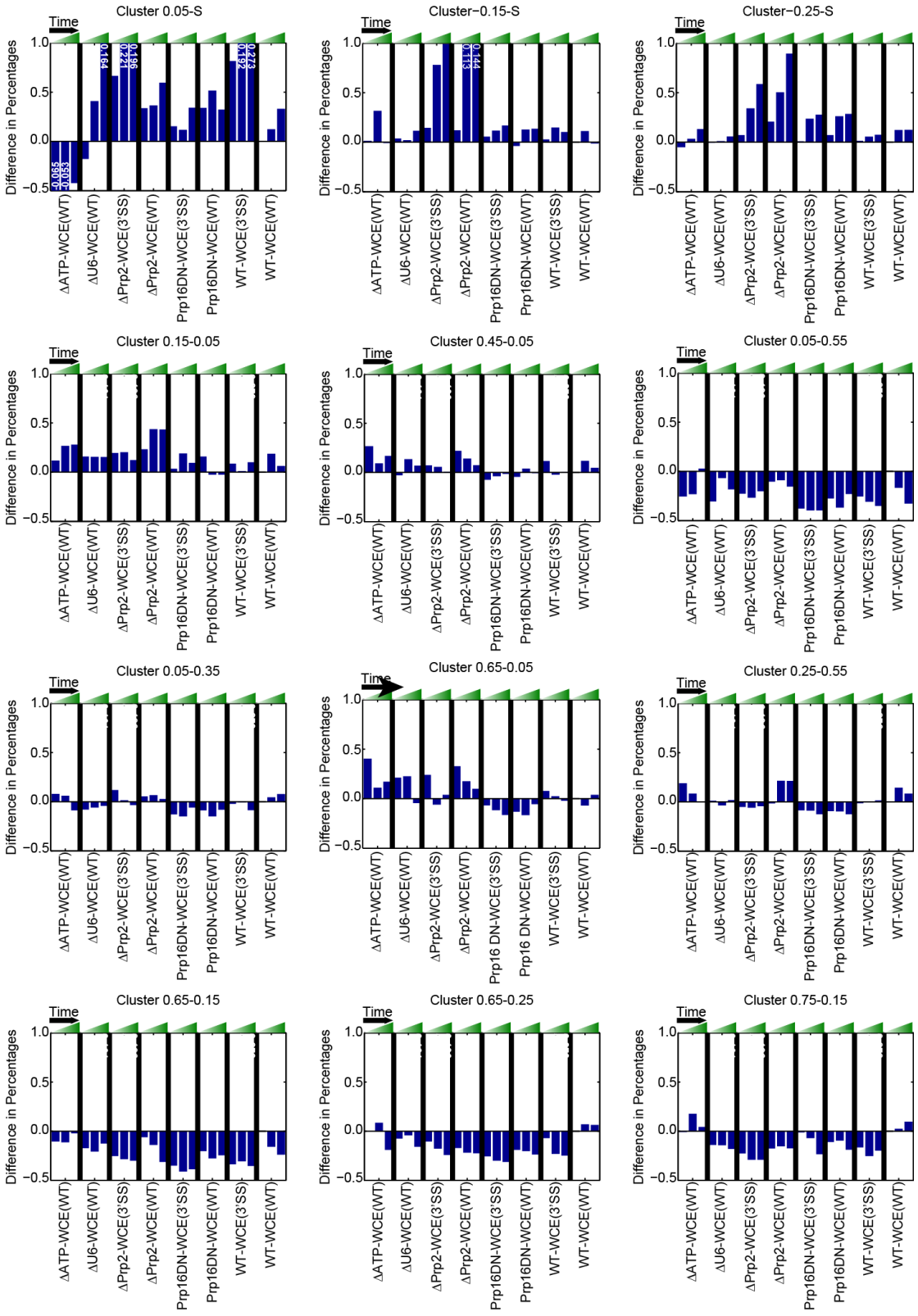
Comparison of the WT with the 3'SS substrate under ΔPrp2-WCE conditions revealed enrichment not only of the 0.05-**S** cluster, characteristic of molecules in an A-like conformation, but also of the 0.65-**S** and 0.75-**S** clusters specifically in the case of the 3'SS mutant (**Figure 3.16**). The latter two clusters are characteristic of the CC2 complex and decrease over time. Previous work on other substrates has suggested that the identity of the 3'SS does not affect assembly of the spliceosome and that recognition and proofreading do not occur until after the first step of splicing[85]. Our results indicate that Ubc4 may behave slightly differently, perhaps due to its altered secondary structured relative to the WT[55]. Deletion of Prp2 may give the spliceosome ample time in the $B^{act}$ stage to detect and discard or reverse-assemble on the 3'SS mutant substrate. Yet, once assembly is allowed to proceed unimpeded past the first step to the C complex, the spliceosome no longer has sufficient time to detect and discard the mutant substrate or reverse-assemble on the substrate. As a result, the subtle differences in assembly become muted and the two substrates behave more similarly (**Figure 3.16**).

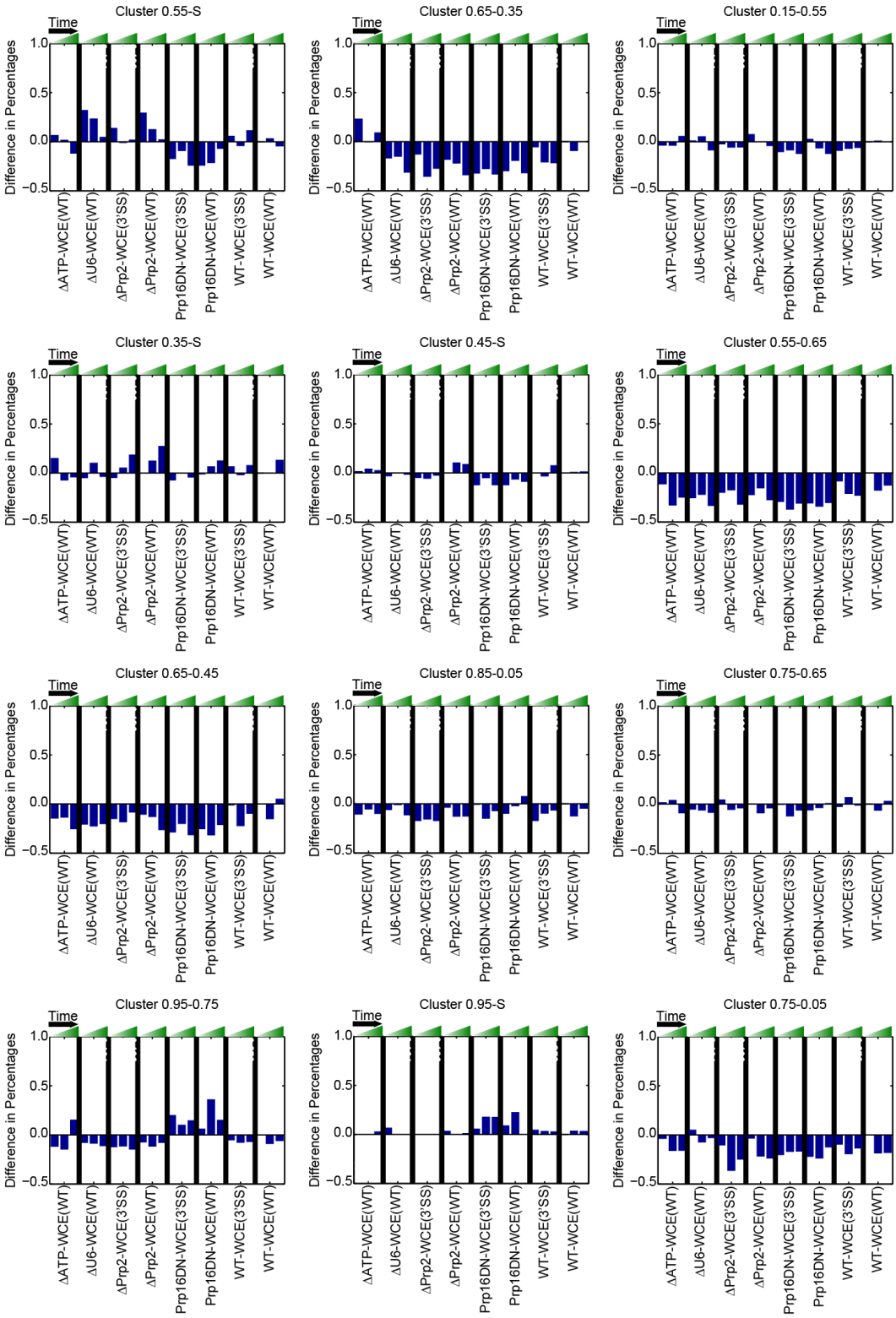
### 3.7 Supplementary Note 3

The initial set of experimentation used for input into SiMCAn involved smFRET data collected solely from incubation either the 3'SS or WT substrate with unmodified extract (WT-

WCE(3'SS) and WT-WCE(WT), **Figure 3.24**). Clustering by SiMCAn resulted in formation of 5 dynamic clusters in addition to the 10 static clusters. Interestingly, cluster 0.05-**S** became increasingly enriched with the 3'SS mutant but remained nearly constant with the WT substrate. This initial set of exploratory research led us to a hypothesis that the 3'SS substrate adopts a static, low FRET conformation after the Bact complex that we wanted to further test. We thus later added in the Prp16DN experiments to determine if this low FRET conformation is formed before or after the first step of splicing.

With the now complete dataset, SiMCAn again identified a 0.05-**S** cluster (clade I) as particularly enriched in the 3'SS mutant substrate after the Prp16-dependent reorganization of the spliceosome. The ATPase Prp16 is known to crosslink to the 3'SS and is required for formation of a functional step 2 active site immediately following the first step of splicing[129]. In addition, one of the second-step splicing factors, Slu7, a protein known to also bind mutant 3'SS, was proposed to be involved with efficient docking of the 3'SS into the step 2 active site[130]. The deficiency in docking observed with the 3'SS may be the result of Slu7 and other second-step factors preventing docking, or the result of the ATPase activity of Prp22. In this latter case, the 3'SS may transiently dock into the second step conformation, but Prp22 rapidly recognizes and discards the mutated 3'SS. Either hypothesis would explain the accumulation of a discarded, undocked substrate unable to proceed through the second step of splicing and also provides further justification for SiMCAn being an excellent form of exploratory analysis capable of providing further hypothesis driven experimentation.


Cluster 0.05-S
Cluster−0.15-S
Cluster−0.25-S
Cluster 0.15-0.05
Cluster 0.45-0.05
Cluster 0.05-0.55
Cluster 0.05-0.35
Cluster 0.65-0.05
Cluster 0.25-0.55
Cluster 0.65-0.15
Cluster 0.65-0.25
Cluster 0.75-0.15
Time
Difference in Percentages
ΔATP-WCE(WT)
ΔU6-WCE(WT)
ΔPrp2-WCE(3'SS)
ΔPrp2-WCE(WT)
Prp16DN-WCE(3'SS)
Prp16DN-WCE(WT)
WT-WCE(3'SS)
WT-WCE(WT)


Cluster 0.55-S
Cluster 0.65-0.35
Cluster 0.15-0.55
Cluster 0.35-S
Cluster 0.45-S
Cluster 0.55-0.65
Cluster 0.65-0.45
Cluster 0.85-0.05
Cluster 0.75-0.65
Cluster 0.95-0.75
Cluster 0.95-S
Cluster 0.75-0.05
Time
Difference in Percentages
1.0
0.5
0.0
−0.5
ΔATP-WCE(WT)
ΔU6-WCE(WT)
ΔPrp2-WCE(3'SS)
ΔPrp2-WCE(WT)
Prp16DN-WCE(3'SS)
Prp16DN-WCE(WT)
WT-WCE(3'SS)
WT-WCE(WT)

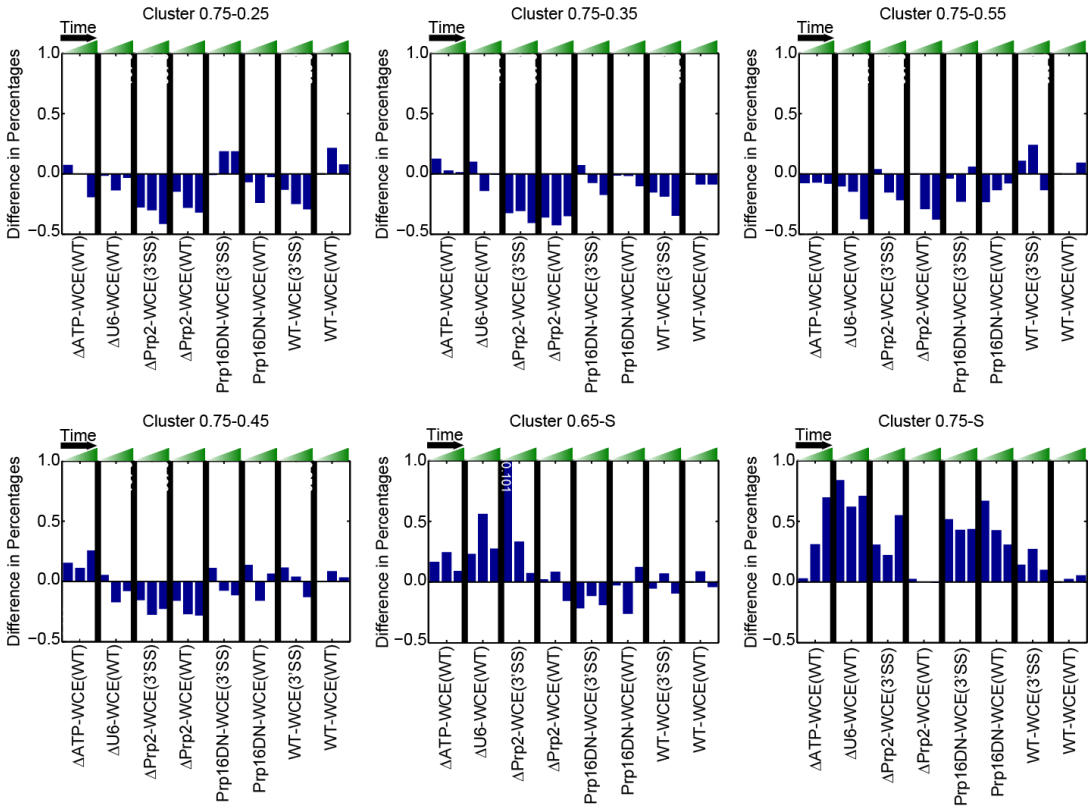



**Figure 3.22 Experimental datasets show vastly different cluster occupancies**
The occupancy of clusters within each of the 8 experimental conditions compared to that of WT-WCE(WT), as in Supplementary figure 13, for the remaining 30 clusters.

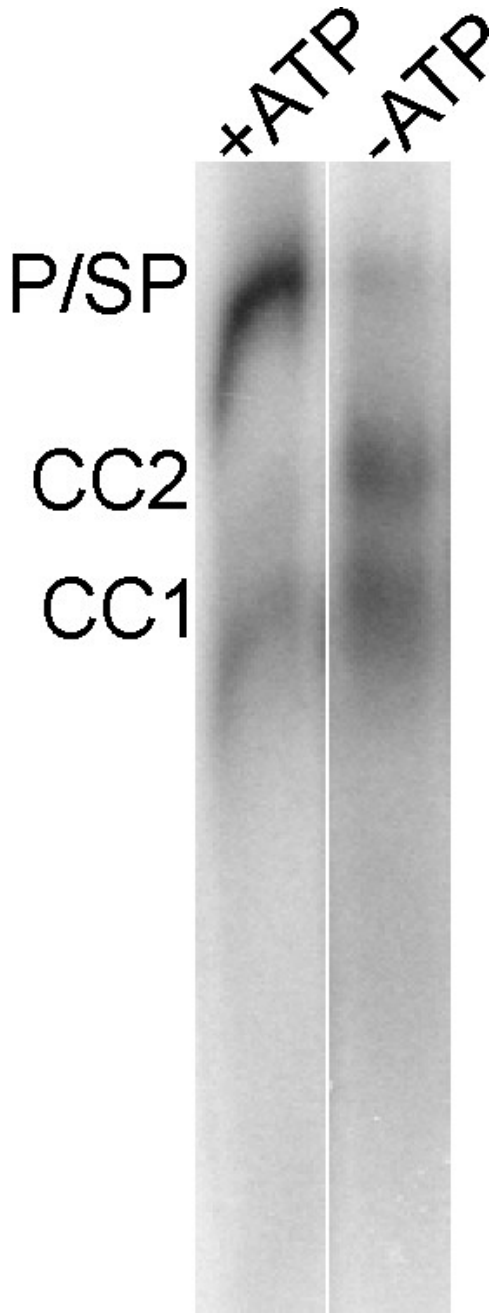



**Figure 3.23 Native gel analysis of commitment complex formation upon Ubc4 in BJ2168 extract**

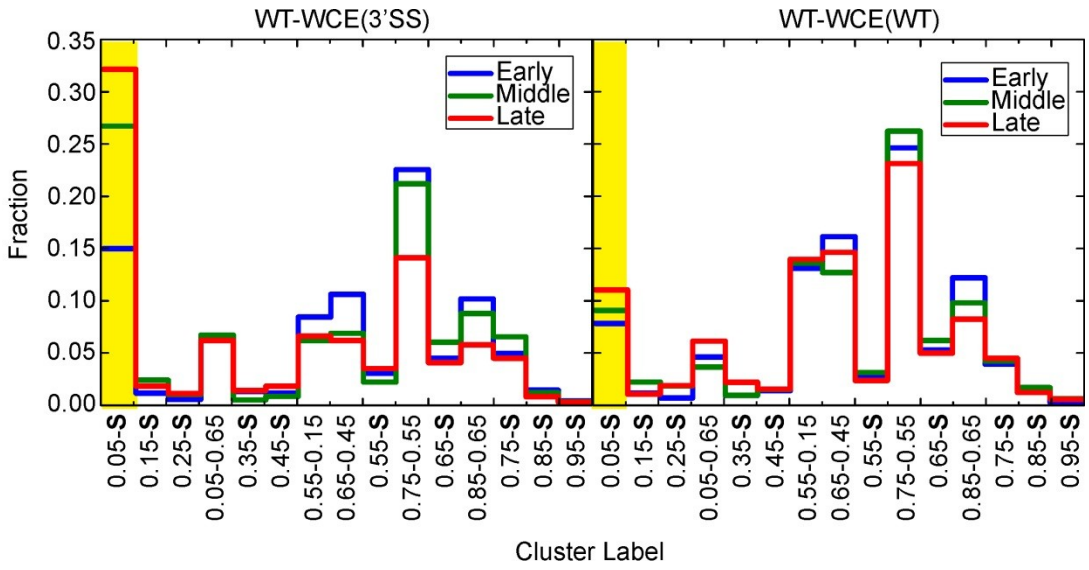



**Figure 3.24 Clustering of substrates in WT extract reveals enrichment of the 0.05-S cluster with the mutant substrate**

Clustering of molecules belonging to the WT-WCE(3'SS) and WT-WCE(WT) conditions reveals enrichment of the 0.05-**S** cluster with the mutant substrate. Clustering resulted in 5 dynamic and 10 static clusters.

**3.8 Acknowledgements**

The authors wish to thank A. Price (University of California, San Francisco) for providing native gel analysis of CC2 formation using Ubc4; D.R. Semlow and J.P. Staley (University of Chicago, Illinois) for providing the dominant negative Prp16 protein expression plasmid; Nguyen N. (Josh) Vo for compiling all MATLAB scripts of SiMCAn into a GUI; as well as C. Guthrie (University of California, San Francisco), D.R. Semlow, J.P. Staley (University of Chicago, Illinois) and A.A. Hoskins (University of Wisconsin, Madison) for providing valuable comments on the manuscript.

# CHAPTER 4: Translating Single-Molecule FRET Traces into the Trajectories of an RNA on its Folding Landscape[3]

## 4.1 Introduction

Proper RNA function relies on the multi-scale conformational dynamics and structure that occur during all RNA processing pathways, often in response to cellular signals. These dynamics vary in their complexity and function, but generally play a key role in every aspect of cellular RNA metabolism, such as in RNA transcription, splicing and translation[111-113]. Whether acting as a standalone RNA molecule, such as with a small ribozyme, or in a large ribonucleoprotein complex, such as with the spliceosome, RNA molecules achieve these diverse biological functions through proper folding into native conformations that properly display specific nucleotides and secondary structure features. As a result, many computational studies focus on predicting a single, native conformation. However, RNA folding can lead to multiple native-like structures, resulting in a great deal of conformational variability (heterogeneity) in the RNA folding landscape. Perhaps the greatest evidence for this comes from the study of self-cleaving ribozymes in which case these alternative conformations were attributed to misfolded, inactive structures. Unfortunately, our understanding of the RNA sequence-structure-function relationship is very limited, and the computational modeling frequently used for RNA structure

[3] Matt Kahlscheuer performed all of the smFRET experimentation and data analysis as well as the biochemical footprinting of the RNA. Nikolai Hecker and Jing Qin at the Center for Non-coding RNA in Technology and Health, University of Copenhagen developed the FRETtranslator software. Peter Kerpedjiev at the Institute for Theoretical Chemistry, University of Vienna assisted in the prediction of RNA 3D sampling structures. A manuscript detailing this material is currently in preparation.

prediction is still in its infancy and often lacks experimental evidence in support of a predicted structure. In recent years, a great deal of work has gone into the development of reliable RNA structure predictions using biochemical footprinting approaches[132,133] in which an RNA modifying reagent is added to an assumed homogenous population of folded RNA. The modifying reagent either cleaves the RNA backbone in accessible/base-paired regions or modifies the RNA backbone/nucleotide, but in both cases is detected as a stall in extension by a polymerase. Too often, however, heterogeneity present in a pool of RNA is overlooked with these techniques due to the fact that they are almost exclusively ensemble approaches that are only capable of capturing a snapshot of the average RNA structure within a solution. As a result, a true understanding of RNA structure prediction based on computational and biochemical footprinting data alone has thus far been insufficient.

Single molecule fluorescence approaches have recently emerged as a powerful toolset to dissect the structure and structural dynamics that form the foundation of biomolecular machines. In particular, single molecule fluorescence resonance energy transfer (smFRET) has been implemented to dissect the RNA dynamics and heterogeneity present in small systems, such as isolated riboswitches[134,135], and large and more complex ones, such as spliceosome assembly and catalysis[53-55,62,110]. Unlike ensemble structure probing, smFRET allows for the observation of time-dependent changes in structure for individual molecules, providing information about subpopulations of behaviors or folding kinetics for a specific RNA. Unfortunately, methods capable of accurately translating smFRET traces into the trajectories of an RNA folding landscape are lacking. As a result, the FRET information of individual pre-mRNA molecules progressing through splicing, for example[62], provides very little structural information for regions of the RNA other than the distance of the specific fluorophore attachment sites.

To address these shortcomings, we have developed the FRETtranslator algorithm, a computational approach capable of incorporating smFRET data for the experimentally supported prediction of RNA secondary structure. We first optimized the algorithm using smFRET data for a 76-nucleotide long RNA containing donor and acceptor dyes in a flexible region of the RNA. We then began to apply the FRETtranslator algorithm to two smFRET data sets from a 135 nucleotide long RNA of which biochemical footprinting data was also gathered. In the future, we plan to use previously analyzed smFRET data[55,62] and known binding sites of several spliceosomal components to better understand the structure of the Ubc4 pre-mRNA within the spliceosome as it progresses through spliceosome assembly.

## 4.2 Materials and Methods

### 4.2.1 Synthesis of truncated and full-length Ubc4 constructs

The truncated, 76-nucletotide long Ubc4 substrate containing Cy3 and Cy5 used in this study was synthesized as previously described (**Table 4.1**)[53,136]. Briefly, the RNA was purchased containing 5-amino-allyl-uridine at position +14 relative to the 5' end and a 5' biotin for immobilization. The RNA was coupled to Cy3 N-hydroxysuccinimidyl ester (GE Healthcare) at the +14 position by resuspending 4 nanomoles of RNA in 40 μL of 0.1 M sodium bicarbonate buffer, pH 9.0, and incubating for 30 min at 60 °C with the dye pack dissolved in DMSO. The conjugated RNA was ethanol precipitated and washed with 70% (v/v) ethanol to remove unconjugated dye. Unlabeled RNA was removed by purification on Benzoylated naphthoylated DEAE (BND)-cellulose (Sigma) that was washed with 1 M NaCl containing 5% (v/v) ethanol. Fully conjugated RNA was eluted with 1.5 M NaCl containing 20% (v/v) ethanol and further precipitated to remove excess salt. The 3' terminus of the RNA was subsequently labeled

| | |
|---|---|
| Full-length UBC4-1 | 5'-GAACUAAGUGAUC(5-N-U)AGAAAGGUAUGUCUAAAGU UAUGGCCACGUUUCAAAUGCGUGCUUUUUUUUUAAAACU UAUGCUCUUAUUUACUAACAAA(5-N-U)CAACAUGCUAUUG AACUAGAUAUCCACCUACUUCAUGUU-3' |
| Full-length UBC4-2 | 5'-GAACUAAGUGAUC(5-N-U)AGAAAGGUAUGUCUAAAGU UAUGGCCACGUUUCAAAUGCGUGCUUUUUUUUUAAAACU UAUGCUCUUAUUUACUAACAAAUCAACAUGCUAUUG AACUAGAGA(5-N-U)CCACCUACUUCAUGUU-3' |
| Truncated UBC4 | 5'-GAACUAAGUGAUC(5-N-U)AGAAAGGUAUGUCUAAAGU UAUGGCCACGUUUCAAAUGCGUGCUUUUUUUUUAAAACU UAUG(Cy5)-3' |
| DNA splint (dSplint) | 5'-GTTGATTTTGTTAGTAAATAAG(SP9)GTTTTAAAAAAAAAGCACGC-3' |

**Table 4.1 Sequence information of oligonucleotides used in this study**
The allyl-amine modified uridines are denoted as (5-N-U). The red and green colors represent positioning of the Cy5 and Cy3 fluorophores, respectively.

through oxidation using sodium periodate and conjugation with hydrazide-containing fluorophore. Briefly, 200 ng of RNA was incubated with 0.1 M sodium periodate in 0.01 M sodium acetate at room temperature for 2 hours in the dark. Precipitate RNA to remove sodium periodate and excess salt. For labeling, RNA was dissolved in 40 uL of 125 mM sodium acetate and combined with 10 µl of hydrazide dye in DMSO and incubated at room temperature for 4 hours. Excess dye was removed through phenol chloroform extraction and ethanol precipitation.

Full length Ubc4 substrates Ubc4-1 and Ubc4-2 containing Cy3 and Cy5 was synthesized and labeled as previously described (**Table 4.1**)[53]. The 135-nucleotide pre-mRNA was ligated from two fragments: a 76-nucleotide 5' segment with 5-amino-allyl-uridine at the -7 position relative to the 5'SS and a 59-nucleotide 3' segment with 5-amino-allyl-uridine at either the +6 or +29 position relative to the BP adenosine (**Table 4.1**). The 5' and 3' fragments were coupled to Cy5 and Cy3 N-hydroxysuccinimidyl ester (GE Healthcare), respectively, by resuspending 4 nanomoles of RNA in 40 µl of 0.1 M sodium bicarbonate buffer, pH 9.0, and incubating for 30 min at 60 °C with the proper dye pack dissolved in DMSO. The conjugated fragments were ethanol precipitated and washed with 70% (v/v) ethanol to remove unconjugated dye. Unlabeled RNA was removed by purification on benzoylated naphthoylated DEAE (BND)-cellulose (Sigma) that was washed with 1 M NaCl containing 5% (v/v) ethanol. Fully labeled RNA fragments were eluted with 1.5 M NaCl containing 20% (v/v) ethanol and further precipitated to remove excess salt. Labeled fragments were combined with an equal molar amount of DNA splint (**Table 4.1**) and ligated by incubating with RNA Ligase 1 (NEB) for 4 h at 37 °C as described[53,117]. Full length, labeled Ubc4 was then purified on a denaturing 7 M urea, 15% (v/v) polyacrylamide gel.

A clone of Ubc4 in pUC19 was used as template for DNA amplification prior to transcription. Purified, double-stranded DNA (ddDNA) was transcribed *in vitro* for 4 h at 37 °C using 2 μg of template in a buffer containing 120 mM HEPES-KOH (pH 7.5), 2 mM spermidine, 30 mM $MgCl_2$, 0.01% (v/v) Triton X-100, 0.1U PPiase, 40 mM DTT, 7.5 mM each NTP, and 0.065 mg/ml homemade T7 RNA polymerase. RNA was purified on a denaturing 7 M urea, 10% (v/v) polyacrylamide gel and stored in water.

*In vitro* transcribed Ubc4 was 5' end labeled by first treating with Antarctic Phosphatase (NEB) for 1 h at 37 °C to remove any 5' phosphate. Enzyme was then inactivated at 70 °C for 5 min, phenol chloroform extracted, and the RNA ethanol precipitated. Labeling was performed with 25 pmoles of RNA in the presence of 20 U polynucleotide kinase (PNK, NEB) and 100 uCi [gamma-$^{32}$P]ATP for 1 h at 37 °C. Labeled RNA was purified on a denaturing 7 M urea, 10% (v/v) polyacrylamide gel and stored in water at 500,000 cpm/ul.

*In vitro* transcribed Ubc4 was 3' end labeled by combining 25 pmoles of RNA with 100 uCi [$^{32}$P]pCp and 10U T4 RNA Ligase (NEB) and incubating at 16 °C for 12-16 h. Unincorporated pCP and degradation products were removed by purification on a denaturing 7 M urea, 10% (v/v) polyacrylamide gel and stored in water at 500,000 cpm/μl.

4.2.2 smFRET analysis of RNA constructs

Single molecule FRET was carried out using a prism-based TIRF microscope[45,49,122], a 532-nm laser to excite the donor (Cy3), and a 635-nm laser to excite the acceptor (Cy5) with the emission recorded at 100 ms time resolution with a Princeton Instruments, I-PentaMAX intensified CCD camera. Biotinylated substrates were heated to 90 °C for 2 min and allowed to cool to RT for at least 10 min in imaging buffer (20 mM Tris-HCl (pH 7.0), 1 M NaCl, 1 mM EDTA). Folded RNA was then immobilized on biotin-PEG coated slides that were pre-reacted with streptavidin,

allowed to bind for 10 min, and then removed of excess RNA by washing with imaging buffer. Data were collected in the presence of imaging buffer containing an oxygen scavenging system (OSS) composed of protocatechuate dioxygenase, protocatechuate and Trolox though directly exciting of Cy3 and recording of Cy3 and Cy5 emission intensities. Histograms were constructed by sampling 100 frames of data from each molecule. The vbFRET software suite[73,123] was used for Hidden Markov Modeling (HMM) in which each trajectory was individually fit with models ranging from 1-5 states with the optimal number of states determined by the vbFRET algorithm.

4.2.3 Terbium(III) footprinting of Ubc4

The structure of $^{32}$P labeled Ubc4 was probed essentially according to the method previously described[47,137-140]. Briefly, a pool of $^{32}$P-Ubc4 with 350,000 c.p.m. of end-labeled RNA per reaction was folded in the presence of 20 mM Tris-HCl (pH 7.0), 1 M NaCl, 0.5 mM EDTA by heating to 90 °C for 90 seconds and snap cooling on ice for 2 min. The folded RNA was then aliquoted into separate tubes and terbium(III)-mediated cleavage was initiated by addition of 1.0 mM $TbCl_3$. Reactions were allowed to proceed for 1 h at 30 °C before addition of 10 mM EDTA to quench the reaction. Cleaved RNA was diluted with 0.3 M NaOAc and precipitated for analysis on a denaturing 7 M urea, 10% (v/v) polyacrylamide sequencing gel. Cleavage product bands were visualized by exposing the gel to a phosphor screen and scanning on a Typhoon variable mode imager (GE Healthcare). The normalized extent of cleavage (Π) was calculated by substituting the peak intensities in the following equation:

$$\prod = \frac{\left( \frac{\text{band intensity at nucleotide } x}{\sum_i \text{band intensity at nucleotide } i} \right)_{y[Tb^{3+}]}}{\left( \frac{\text{band intensity at nucleotide } x}{\sum_i \text{band intensity at nucleotide } i} \right)_{0\,\text{mM}\,[Tb^{3+}]}}$$

where y is the terbium(III) concentration in a particular cleavage reaction and x is the nucleotide position of the RNA.

4.2.4 Computational transformation of smFRET data into structural folding pathway

The FRETtranslator algorithm requires an RNA sequence and smFRET trace as input to predict the most likely transitions between RNA secondary structures based on the Viterbi decoding algorithm, a dynamic programming algorithm for finding the most likely sequence of hidden states (**Figure 4.1**)[141]. The Viterbi algorithm requires 5 input parameters (the hidden states, transition rates, emission rates, and emission probabilities), the first three of which can be described using the recently developed Basin Hopping Graph (BHG)[142]. The BHG is a coarse-grained model of the folding landscape for an RNA sequence where each node of the BHG is a local minimum that represents a basin in the landscape (a hidden state). Such local minima are computed from the set of all possible secondary structures that can be formed assuming that only canonical (GC, AU, GU) base pairs are formed, base pairs do not cross (no pseudoknots), and hairpin loops have a minimum length of three. The basins are then arranged as a graph by next creating the "move set" in which it is determined if specific secondary structures can be interconverted in a single step. The BHG edges connect basins when the direct transitions between them are energetically favorable with transition rates being based on the barrier heights within the BHG. The initial probability is the Boltzmann distribution of the free energy for each

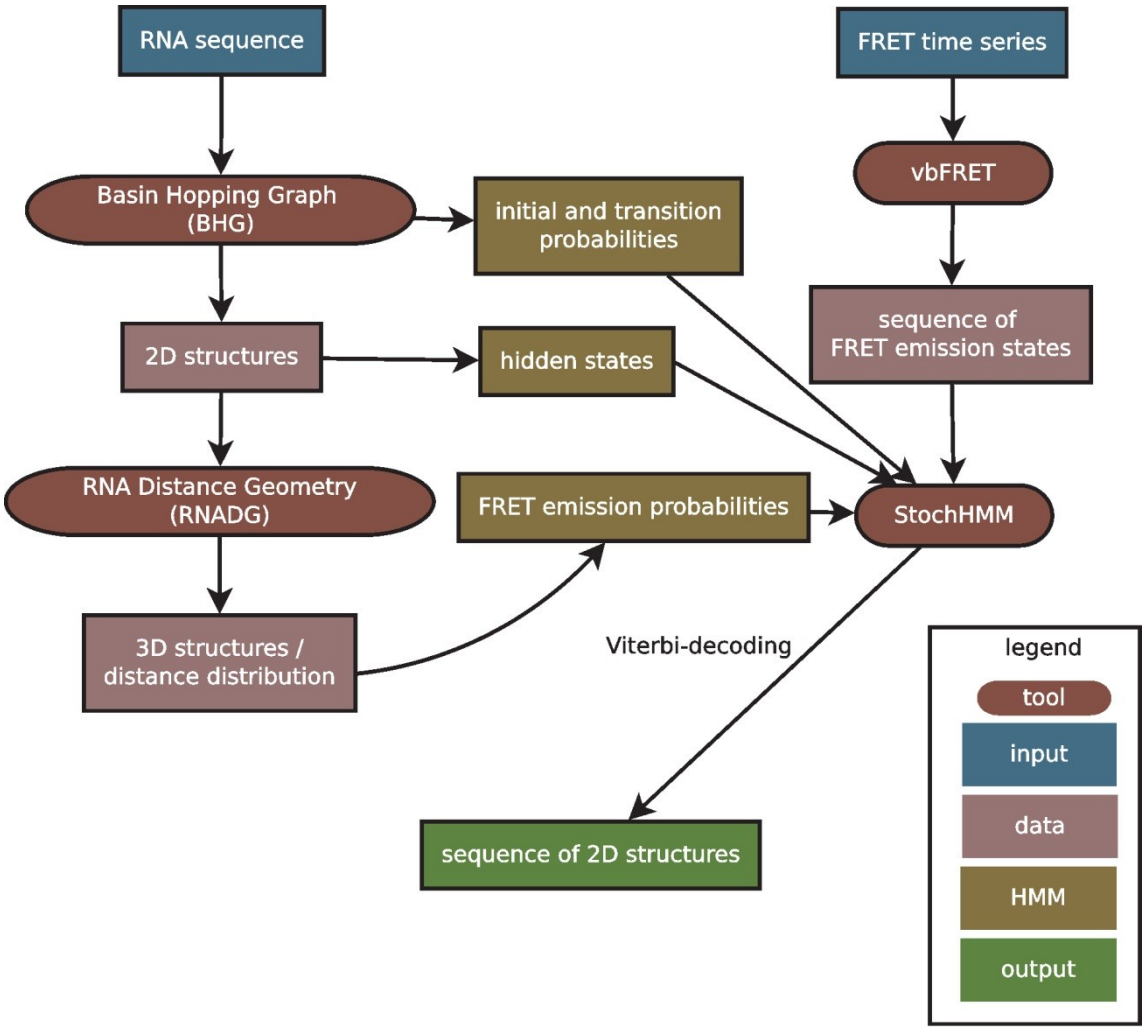



**Figure 4.1 Illustration of the computational framework of FRETtranslator**

local minima of the BHG folding landscape. The sequence of FRET state observations for each single-molecule trajectory becomes the emission rates. Lastly, emission probability distributions are determined for each hidden state secondary structure. Several hundred predicted 3D structures are determined for each secondary structure for which the Euclidean distance between the atoms closest to the two fluorophore positions in sampled 3D space is determined yielding a distribution of distances for each secondary structure. Based on the distribution of distances, FRET efficiency values (E) can be calculated to yield a distribution of FRET states for each secondary structure using the Förster equation, $E = \left(1 + \left(\frac{R}{R_0}\right)^6\right)^{-1}$, where $R_0$ is the Förster radius for the Cy3-Cy5 FRET pair at which their FRET efficiency is 50% (54 Å) and R is the donor-acceptor distance as described[74,76,77]. The idealized FRET trace from vbFRET is then combined with the emission, initial, and transition probabilities and used as input for Viterbi-decoding. Finally, the Viterbi path yields the most-likely sequence of secondary structures.

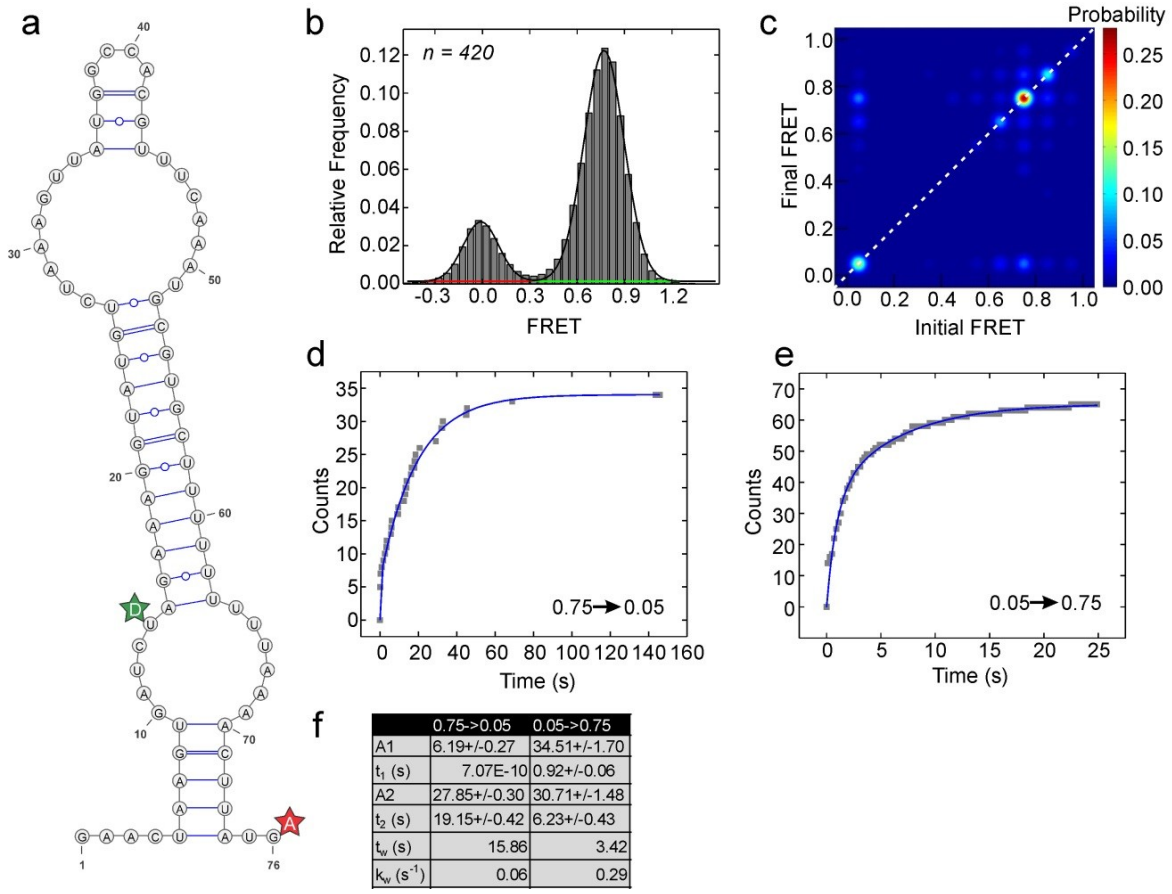
## 4.3 Results

### 4.3.1 Designing a suitable smFRET RNA substrate

Rapid computational modeling of RNA secondary structures has greatly improved with the application of the nearest neighbor model to explain the thermal melting data for an RNA molecule[143]. It is well known that the number of predicted structures grows exponentially with the length of the RNA sequence due to the introduction of pseudoknots and other tertiary structural features which make the prediction process more difficult and time-consuming[144]. Unfortunately, there is little experimental data available with which to refine or confirm the computational predictions. In addition, the majority of RNA folding investigations treat RNA folding as a stochastic process that is defined by an RNA folding landscape containing one stable

structure. RNA folding, however, is known to be a dynamic process with often no one stable structure but many interconverting structures of similar energies in its folding landscape. Single-molecule fluorescence resonance energy transfer (smFRET) is a biophysical technique capable of producing a time series of FRET signals and rates of exchange between them and can thus identify subpopulations of behaviors and structures that would become averaged in most biochemical structure analysis techniques[45,73,122]. We thus set out to improve current models for the pre-mRNA structure within the spliceosome by integrating smFRET data into RNA secondary structure predictions. In a subsequent step, this will allow us to refine current RNA 3D structure models for the spliceosome.

Although the BHG algorithm has greatly enhanced our ability to sample relevant and sufficiently diverse secondary structures for RNAs that are 100s of nucleotides in length, 3D structure modeling of RNA beyond 100 nucleotides is time-consuming and inaccurate. The Ubc4 pre-mRNA that has been previously characterized by smFRET throughout spliceosome assembly and catalysis has been shortened to allow investigation by smFRET and still be an efficient splicing substrate. Unfortunately, the resulting pre-mRNA substrate is still well over 130 nt in length and just beyond the capabilities of current 3D structure prediction. We therefore first synthesized and labeled a further truncated form of Ubc4 only 76-nts in length that would be more suitable for development and optimization of the FRETtranslator algorithm. Secondary structure prediction using the mFOLD web server[145] revealed one dominant structure with a free energy of -14.4 kJ/mol (**Figure 4.2a**). This construct was synthesized containing a 5' biotin for immobilization and an allyl-amine modified uridine residue for attachment of the Cy3 fluorophore, while the Cy5 acceptor fluorophore was attached to the free 3' end of the RNA using oxidation chemistry.




f

| | 0.75->0.05 | 0.05->0.75 |
|---|---|---|
| A1 | 6.19+/-0.27 | 34.51+/-1.70 |
| $t_1$ (s) | 7.07E-10 | 0.92+/-0.06 |
| A2 | 27.85+/-0.30 | 30.71+/-1.48 |
| $t_2$ (s) | 19.15+/-0.42 | 6.23+/-0.43 |
| $t_w$ (s) | 15.86 | 3.42 |
| $k_w$ ($s^{-1}$) | 0.06 | 0.29 |

**Figure 4.2 smFRET analysis of the further truncated Ubc4 substrate**
(a) The truncated Ubc4 substrate contains donor (green, Cy3) and acceptor (red, Cy5) dyes at the +14 and 3' end positions, respectively. The structure shown is the lowest energy structure as predicted using mFOLD. (b) FRET probability distribution of the raw single-molecule FRET trajectories. (c) TODP showing the fraction of molecules that either do not show a transition (dotted white line) or transition from one indicated FRET state to another. (d and e) Cumulative distribution plot of dwell times extracted for the indicated transition and fit with a double-exponential rate equation (f) Parameters for the double-exponential equations fitted to the dwell time data. Weighted average rate constants ($k_w$) were calculated by utilizing the amplitudes associated with each time constant as weighting factors.

4.3.2 Truncated Ubc4 adopts primarily a high FRET conformation

To obtain real-time insight into the structure and conformational rearrangements of the folded, truncated Ubc4 RNA at equilibrium, we immobilized the substrate and carried out smFRET in a high salt buffer at neutral pH. RNA molecules were verified of containing both Cy3 and Cy5 fluorophores prior to collection of Cy3 and Cy5 intensities over 3,000 video frames (at 100-ms time resolution) from 420 molecules. Histogram analysis of the FRET values indicated a bimodal Gaussian distribution with a dominant high FRET peak at 0.77±0.001 in ~82% of molecules and an unfolded structure centered at -0.01±0.005 in ~18% of molecules (**Figure 4.2b**). Hidden Markov model (HMM) analysis is well suited for smFRET analysis because of its ability to find discrete states within noisy time series data and reliably find the most probable path through these smFRET states. We therefore used the freely available software vbFRET[123] to idealize the FRET states and series of states for each individual molecule. To enable a direct comparison across a large dataset, we binned each FRET state into one of ten evenly spaced FRET values (0.05-0.95, with increments of 0.10) that together evenly span the viable FRET range and are commensurate with typical signal-to-noise ratios. Transition occupancy density plots (TODPs), which are scaled to emphasize the most common transitions within a population[73], indicate that the most common behavior is a static, unchanging high FRET state near 0.75 FRET and to a lesser extent, a static low FRET state around 0.05 FRET (**Figure 4.2c,** diagonal molecules are static). Additionally, a small fraction of molecules appear to show transitions between the low and high FRET states (**Figure 4.2c**). In order to identify the rate of transition between the high (~0.75) and low (0.05) FRET states, we performed kinetic analysis by building cumulative distribution plots of the dwell time data for each transition and fitting the data to an exponential equation. Interestingly, the low-to-high FRET transition appears to contain two similarly sized

populations, one fast and one slow transitioning ($t_1$=0.92 and $t_2$=6.23), while the high-to-low transition is dominated by a slow-transitioning population of molecules indicating that the high FRET state is the more stable conformation. Weighted average rate constants reveal that the low-to-high FRET rate of transition is nearly 5-times faster and occurs much more frequently than the high-to-low transition rate (**Figure 4.2d,e,f**), demonstrating that although molecules are able to occupy an unfolded, low FRET state, they tend to quickly transition back to the much more stable high FRET conformation.

4.3.3 FRETtranslator identifies folding trajectories of secondary structures

The vbFRET-idealized smFRET trajectories for the short 76-nt long RNA substrate were next used as input for the FRETtranslator algorithm in order to identify potential 2D structures that could represent the observed high and low FRET states. FRETtranslator found ~2,900 possible secondary structures for each of which ~500 3D structures were predicted. More precisely, we used a distance geometry model that restrains Watson-Crick base pairs and A-RNA helices in typical geometry as input for the molecular modeling package TINKER to sample 3d structures[146-148]. FRETtranslator calculated the Euclidean distance between Cy3 and Cy5 at nucleotides 14 and 76, respectively, for each 3D structure, and converted the distances to FRET efficiencies to yield a distribution of FRET values for each 3D structure. These FRET probability distributions were then compared to the entire smFRET dataset for truncated Ubc4 in order to determine which structures best match each individual FRET trajectory. Finally, predicted structures were analyzed to determine how often a structure is assigned to a given FRET value (**Table 4.2,** top two structures shown per FRET state), as well as how often a predicted structure undergoes a 'self' (transitions to the same structure) or 'non-self' (transitions to another structure) transition (**Table 4.3**).

| FRET/id | 514 | 516 | 568 | 1732 | 2779 | 3115 | 6438 | 7401 |
|---|---|---|---|---|---|---|---|---|
| 0.05 | 0 | 0 | 0 | 0 | 0 | 212 | 0 | 41567 |
| 0.65 | 29628 | 1692 | 0 | 530 | 25 | 1 | 435 | 0 |
| 0.75 | 0 | 95 | 0 | 82921 | 11368 | 0 | 2609 | 5 |
| 0.85 | 0 | 27 | 3692 | 363 | 0 | 0 | 42570 | 0 |

**Table 4.2 The top two most commonly predicted structures for each FRET state**
Shown is the number of time steps a structure ID is predicted for each of the indicated FRET states. Red and green values indicate the first and second most common structure, respectively, for each FRET value.

| trans | count | self | FRET Value |
|---|---|---|---|
| 1732->1732 | 83809 | 1 | 0.75 |
| 6438->6438 | 45535 | 1 | 0.85 |
| 7401->7401 | 41488 | 1 | 0.05 |
| 514->514 | 29655 | 1 | 0.65 |
| 2779->2779 | 11369 | 1 | 0.75 |
| 4615->4615 | 5470 | 1 | 0.75 |
| 341->341 | 4336 | 1 | 0.95 |
| 297->297 | 3739 | 1 | 0.55 |
| 568->568 | 3730 | 1 | 0.85 |
| 6110->6110 | 2544 | 1 | 0.95 |
| 2779->3115 | 40 | 0 | 0.75 --> 0.05 |
| 3115->2779 | 39 | 0 | 0.05 --> 0.75 |
| 3115->7401 | 16 | 0 | 0.05 --> 0.05 |
| 7401->3115 | 14 | 0 | 0.05 --> 0.05 |

**Table 4.3 The top 10 'self' transitions and top 4 'non-self' transitions**
Counts indicate the number of time steps each transition occurs. The FRET value column indicates the FRET value most frequently associated with the indicated structure ID.

Perhaps not surprisingly, the most frequently predicted structure and transition was found to be the 'self' transition of structure 1732 (**Table 4.3**) with nearly 98% of these 'self' transitions assigned to molecules possessing a 0.75 FRET state (**Figure 4.3a**). These data match well with the previous histogram and TODP analysis that showed a static 0.75 FRET state as the most common FRET state and transition (**Figure 4.2b,c**). The second most predicted structure and transition was the 'self' transition observed for structure 6438 (**Table 4.3**) with nearly 93% of the 'self' transitions assigned to molecules possessing a 0.85 FRET state (**Figure 4.3a**). Interestingly, this structure shows great similarity to structure 1732 describing the 0.75 FRET state, specifically in the very stable 5' stem structure where the fluorophores are located. The same 5' stem including dangling ends was also predicted by mFOLD. However, the mFOLD MFE structure differs in the size of the interior loops adjacent to the stem loop. Finally, the third most predicted structure and transition was found to be the 'self' transition of structure 7401 (**Table 4.3**), a highly linear and unfolded form of the RNA that would allow for great separation of the donor and acceptor fluorophores and thus result in the formation of a low, 0.05 FRET state (**Figure 4.3b**), again matching the previous histogram and TODP analysis.

Analysis of the most common 'non-self' transitions revealed two primary sets of transitions (**Table 4.3**), one between the second most predicted 0.75 FRET structure (**Table 4.2**, structure 2779) and the second most predicted 0.05 FRET structure (**Table 4.2**, structure 3115), and the other between the top two 0.05 FRET structures (**Table 4.2**, structures 3115 and 7401). Notably, these two most common sets of transitions (2779←→3115 and 3115←→7401) explain the occurrence of two populations of molecules transitioning from 0.05 to 0.75 FRET in the kinetic analysis (**Figure 4.2e,f**). One population of molecules transitions very rapidly from structure 3115 back to structure 2779, spending on average ~1.4 sec in the low FRET structure

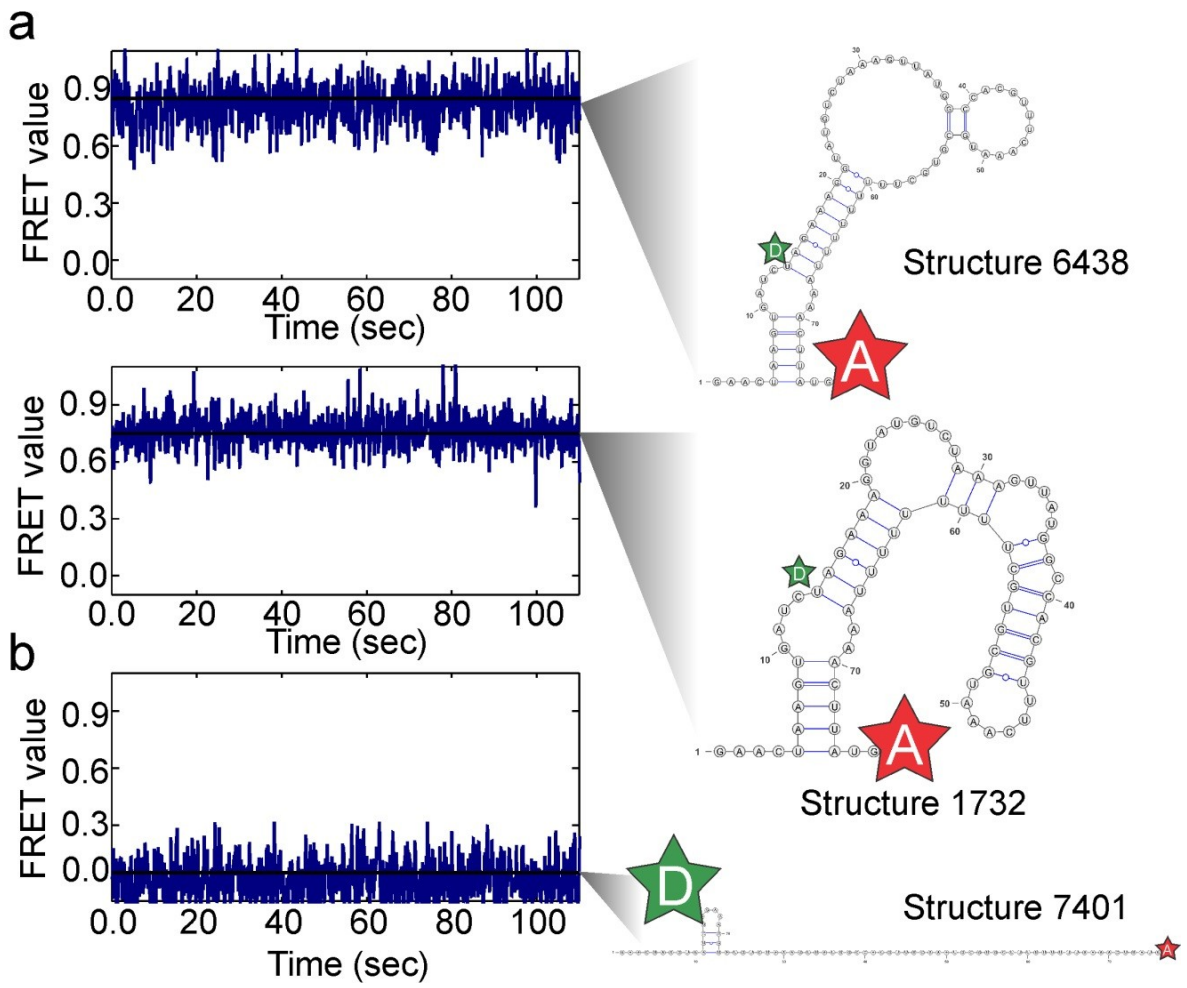



**Figure 4.3 The most common structures predicted by FRETtranslator are in stable high or low FRET states**

Example raw FRET trace (blue), idealized FRET trajectory (black), and corresponding most predicted structure for the three most common FRET states. (a) Structures 1732 and 6438 describe molecules trapped in a static 0.75 or 0.85 high FRET state, respectively. (b) Structure 7401 best describes molecules that exhibit long dwell times in the 0.05 low FRET state.

(**Figure 4.4a**). Conversely, an equally sized population of molecules that transitions from 2779 to 3115 can transition very slowly back to the more stable high FRET conformation, but to do so must further unfold from structure 3115 to structure 7401. Once in the conformation of structure 7401, molecules are able to stably reside in an unfolded conformation for an average of ~14 sec (**Figure 4.4b**). These data match the kinetic analysis of FRET states (**Figure 4.2c,e,f**) that found a population of molecules that transitions very slowly from the low FRET state back to the high FRET state ($t_1$ = 0.92 s) and an equally sized population with longer dwell times in the low FRET state ($t_2$ = 6.23 s). FRETtranslator has, therefore, identified three primary structures describing the high FRET state and two primary structures describing the low FRET state. Structures 1732 and 6438 contain nearly identical 5' stems that result in a stable 0.75 or 0.85 FRET state, respectively. Structure 2779 on the other hand is an alternative and less stable structure describing the high FRET behavior of truncated Ubc4 that is capable of more easily unfolding into structure 3115 possessing a 0.05 FRET state. Once unfolded, this low FRET structure can either very rapidly transition back to the high FRET 2779 structure or continue to unfold to structure 7401, a more stable unfolded form of truncated Ubc4 (**Figure 4.4c**).

4.3.4 Full-length Ubc4 FRET distribution shows high and low FRET behaviors

Having shown that the FRETtranslator algorithm yields reasonable secondary structure predictions for a short RNA substrate, we next sought to apply the same workflow to a biologically more relevant RNA. Ubc4 is a yeast pre-mRNA substrate that has been modified to be suitable for use in single molecule investigation of the mechanism of pre-mRNA splicing[53,55,62]. The 135 nt long pre-mRNA contains a short intron known to possess extensive secondary structure that place the splice sites in close proximity even in the absence of the spliceosome[62,73]. Because of Ubc4's increased length, input into mFOLD for

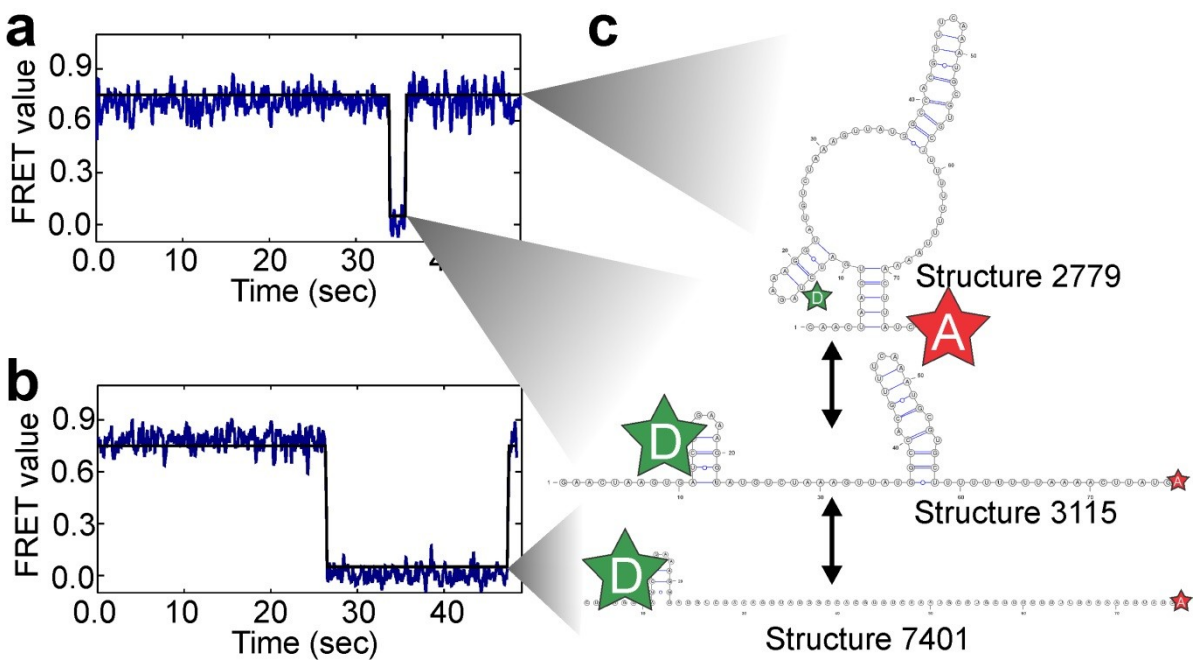



**Figure 4.4 Predicted substrate unfolding to an unstable or stable low FRET state**
(a) An example raw FRET (blue) and idealized FRET (black) trace for a molecule that transitions from a stable high FRET conformation (structure 2779) to an unstable low FRET conformation (structure 3115). (b) An example raw FRET (blue) and idealized FRET (black) trace for a molecule that transitions from a stable high FRET conformation (structure 2779) to a stable low FRET conformation (structure 7401). (c) FRETtranslator-predicted transition from a stable high FRET state to either an unstable low FRET state (3115) or to a stable low FRET conformation (structure 7401).

preliminary secondary structure analysis yields 8 predicted structures with very similar free energies (**Figure 4.5**). To more confidently predict the most stable structures and observe real-time structural interconversions, we synthesized two Ubc4 constructs with different labeling positions of the Cy3 fluorophore. Ubc4-1 contains the Cy3 and Cy5 fluorophores at the +96 and +14 positions (**Table 4.1**), regions that had previously allowed for the observation of branchsite docking into the 5'SS during the 1st step of splicing[55,62]. The fluorophores are also near the base of a helical stem thought to undergo extensive rearrangements even in the absence of the spliceosome and would be predicted to provide valuable structure information for this unstable region of the substrate. The Ubc4-2 construct, on the other hand, has the Cy3 fluorophore positioned at residue +119 (**Table 4.1**), allowing for the surveillance of 5'SS and 3'SS proximity during the 2nd step of splicing[53]. This region of the RNA is thought to be much more stable based on structure prediction and should thus serve as a good reference of correct structure prediction.

Each substrate was individually immobilized for smFRET analysis in the same buffer conditions and experimental setup as with the truncated Ubc4 construct. Histogram analysis of the FRET values for the Ubc4-1 construct indicated a double Gaussian distribution with a high FRET population at 0.69±0.002 for ~57% of the molecules and a low 0.15±0.003 FRET state in ~43% of molecules, indicating that a much larger fraction of molecules visit an alternative confirmation in which the fluorophores become greatly removed from one another (**Figure 4.6b**). TODP analysis revealed a much smaller fraction of molecules exhibiting static behavior. Rather, the dominant behavior is a transition between the numerous high FRET state confirmations and several low FRET states (**Figure 4.6c**). In order to gain an approximate understanding of the transition kinetics between the high and low FRET populations, the 0.65 and 0.75 FRET states were grouped together, as were the 0.15 and 0.25 FRET states, and dwell

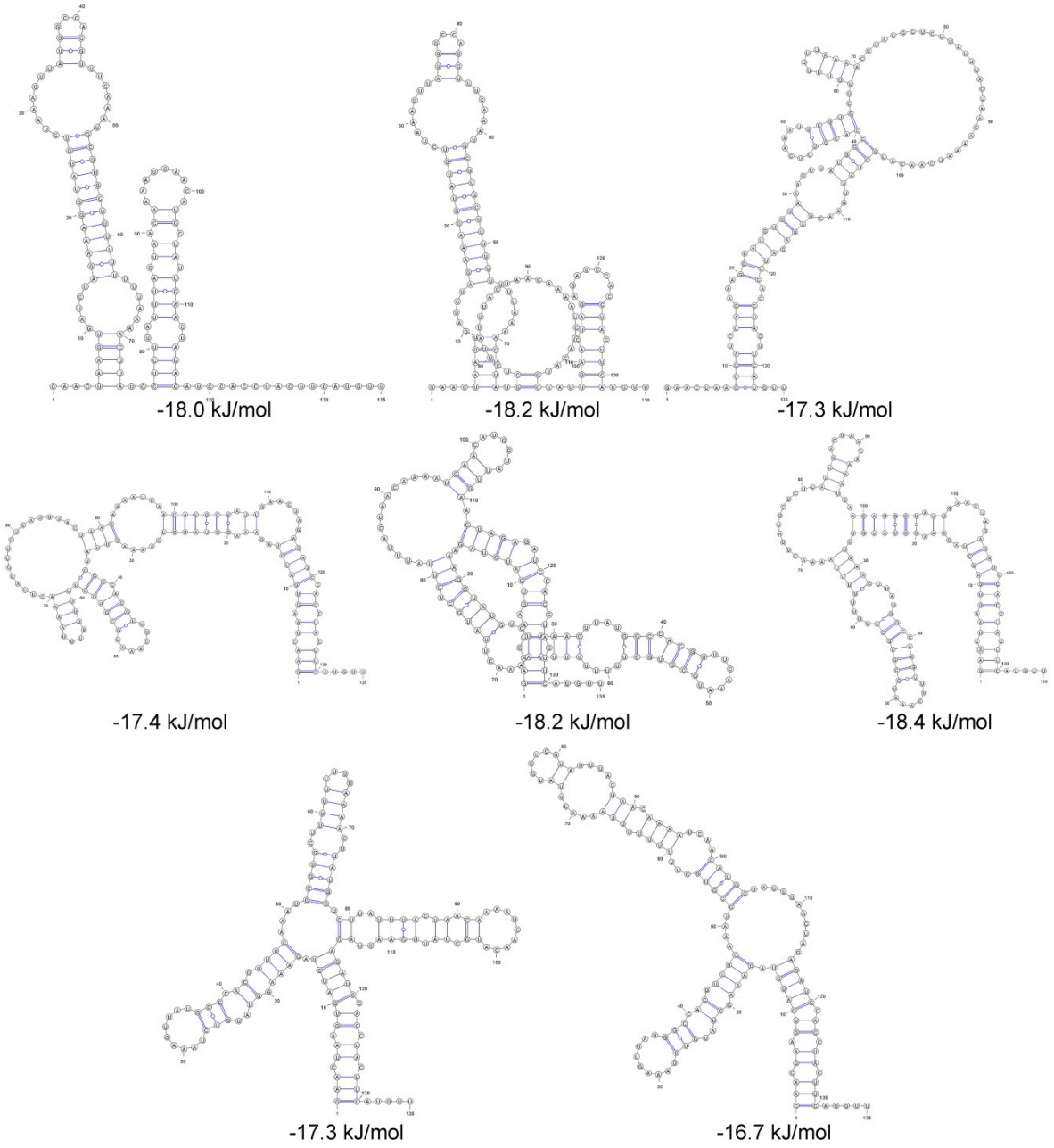
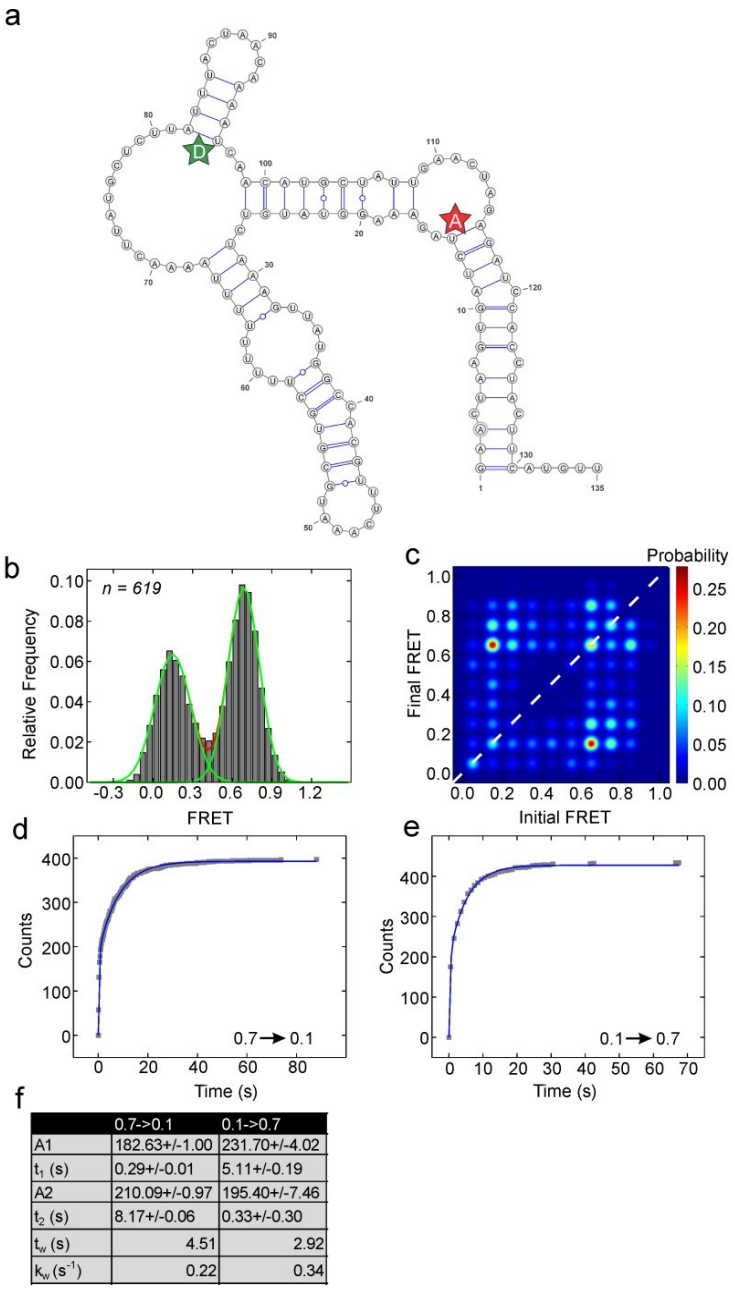



**Figure 4.5 The most stable predicted full-length Ubc4 structures**
The top 8 most stable predicted structures found using mFOLD.




f

| | 0.7->0.1 | 0.1->0.7 |
|---|---|---|
| A1 | 182.63+/-1.00 | 231.70+/-4.02 |
| $t_1$ (s) | 0.29+/-0.01 | 5.11+/-0.19 |
| A2 | 210.09+/-0.97 | 195.40+/-7.46 |
| $t_2$ (s) | 8.17+/-0.06 | 0.33+/-0.30 |
| $t_w$ (s) | 4.51 | 2.92 |
| $k_w$ ($s^{-1}$) | 0.22 | 0.34 |

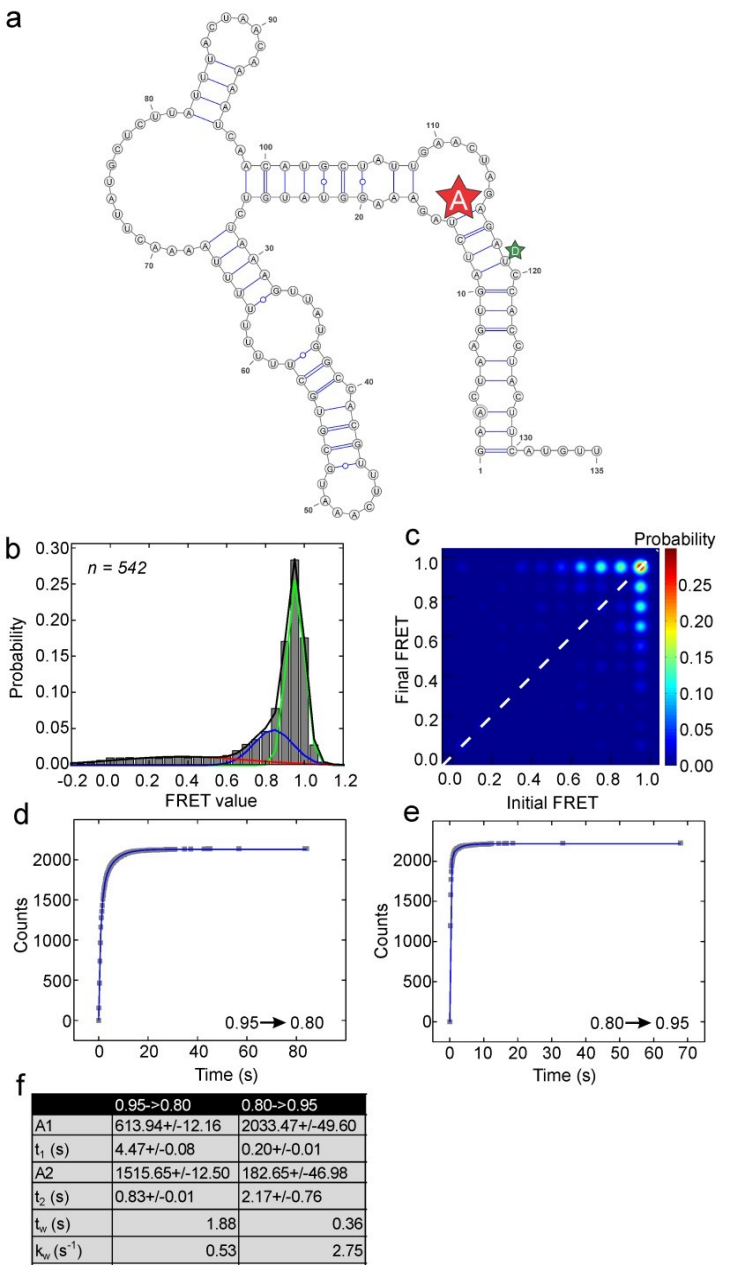
**Figure 4.6 smFRET analysis of full length Ubc4 substrate**
(a) Full length Ubc4-1 contains donor (green, Cy3) and acceptor (red, Cy5) dyes at the +14 and +96 positions, respectively. The structure shown is the lowest energy structure as predicted using mFOLD. (b) FRET probability distribution of the raw single-molecule FRET trajectories. (c) TODP showing the fraction of molecules that either do not show a transition (dotted white line) or transition from one indicated FRET state to another. (d and e) Cumulative distribution plot of dwell times extracted for the indicated transition and fit with a double-exponential rate equation (f) Parameters for the double-exponential equations fitted to the dwell time data. Weighted average rate constants ($k_w$) were calculated by utilizing the amplitudes associated with each time constant as weighting factors.

times analysis was performed. The resulting low-to-high FRET transition rate constant was found to be only slightly higher than the high-to-low FRET transition, indicating that, although molecules appear to be primarily dynamic in nature, they slightly prefer a more compact structure that places the regions of RNA containing the Cy3 and Cy5 fluorophores in close proximity (**Figure 4.6c,d,e**).

As expected, interpretation of the Ubc4-2 smFRET data was much more straightforward, with the histogram and TODP analysis supporting that this region of the RNA adopts primarily a stable, high FRET conformation. The FRET histogram of the data was fit best using a trimodal Gaussian, with a dominant ~57% population possessing a mean high FRET conformation of 0.95 FRET as well as a lesser ~24% population with a mean FRET value of 0.84±0.02 (**Figure 4.7b**). Molecules have a tendency to remain in a static 0.95 FRET state, as indicated by the high probability on the diagonal 0.95 FRET, but occasionally make very fast transitions to a slightly lower 0.8 FRET state (**Figure 4.7c**). Interestingly, kinetic analysis shows that if molecules do transition out of the high FRET state to the 0.8 FRET state they very rapidly transition back to high FRET at a rate 5 times faster than the 0.95→0.80 transition (**Figure 4.7d,e,f**). These data would appear to indicate that the helix formed between these two regions of RNA where the FRET probes are incorporated is highly stable and unlikely to change in FRETtranslator secondary structure predictions.

4.3.5 Biochemical footprinting of Ubc4 reveals single-stranded regions of RNA

To provide further experimental data for input into the FRETtranslator algorithm, we performed terbium(III) footprinting with the folded full-length Ubc4 substrate. Terbium(III) is a metal ion that, at high concentrations, binds and cleaves RNA in single-stranded, flexible regions[139,140]. Biochemical footprinting, such as this, is an ensemble technique that can only provide an average




f

| | 0.95->0.80 | 0.80->0.95 |
|---|---|---|
| A1 | 613.94+/-12.16 | 2033.47+/-49.60 |
| $t_1$ (s) | 4.47+/-0.08 | 0.20+/-0.01 |
| A2 | 1515.65+/-12.50 | 182.65+/-46.98 |
| $t_2$ (s) | 0.83+/-0.01 | 2.17+/-0.76 |
| $t_w$ (s) | 1.88 | 0.36 |
| $k_w$ ($s^{-1}$) | 0.53 | 2.75 |

**Figure 4.7 smFRET analysis of full length Ubc4-2 substrate**
(a) Full length Ubc4-2 contains donor (green, Cy3) and acceptor (red, Cy5) dyes at the +14 and +119 positions, respectively. The structure shown is the lowest energy structure as predicted using mFOLD. (b) FRET probability distribution of the raw single-molecule FRET trajectories. (c) TODP showing the fraction of molecules that either do not show a transition (dotted white line) or transition from one indicated FRET state to another. (d and e) Cumulative distribution plot of dwell times extracted for the indicated transition and fit with a double-exponential rate equation (f) Parameters for the double-exponential equations fitted to the dwell time data. Weighted average rate constants ($k_w$) were calculated by utilizing the amplitudes associated with each time constant as weighting factors.

protection pattern of all interconverting structures within a pool of RNA. However, it does provide nucleotide-resolution protection patterns that, depending upon the degree of protection, can provide a high level of confidence for specific nucleotides being single-stranded or base-paired.

The full-length, *in vitro* transcribed and $^{32}$P-labeled Ubc4 was melted and re-folded in the same high-salt imaging buffer used during smFRET analysis of the fluorescent substrates. Ubc4 was radioactively labeled on either the 5' or 3' end in order to gain a confident footprint for the entire RNA that could be used to further refine the predicted structures. The concentration of terbium(III) and incubation time was optimized and found to ideally be 1.0 mM terbium ion for 1 h at 30 °C resulting in significant cleavage of the RNA backbone (**Figure 4.8a,b**). At least 3 replicates were performed with each substrate and the normalized reactivity for each nucleotide position was determined. In **Figure 4.8c** nucleotides with low reactivity are highlighted in blue, with medium reactivity in yellow, and with high reactivity in red. This analysis revealed 11 nucleotides that show high levels of reactivity with terbium(III) and thus can be predicted with confidence to be single-stranded. When re-inserted into mFOLD for structure determination forcing the nucleotides of high reactivity to be single-stranded now produces two primary structures with similar energy that differ in the length of the L4 stem (**Figure 4.8d**).

**4.4 Discussion**

Here we have combined single-molecule FRET and computational analysis techniques as input for FRETtranslator, an algorithm capable of utilizing smFRET and 3D structure predictions to predict a time-series of RNA structures. An approach such as FRETtranslator allows for the prediction of secondary structures for any RNA of which there are smFRET data available. Predicted secondary structure transitions can be used as input for computationally more

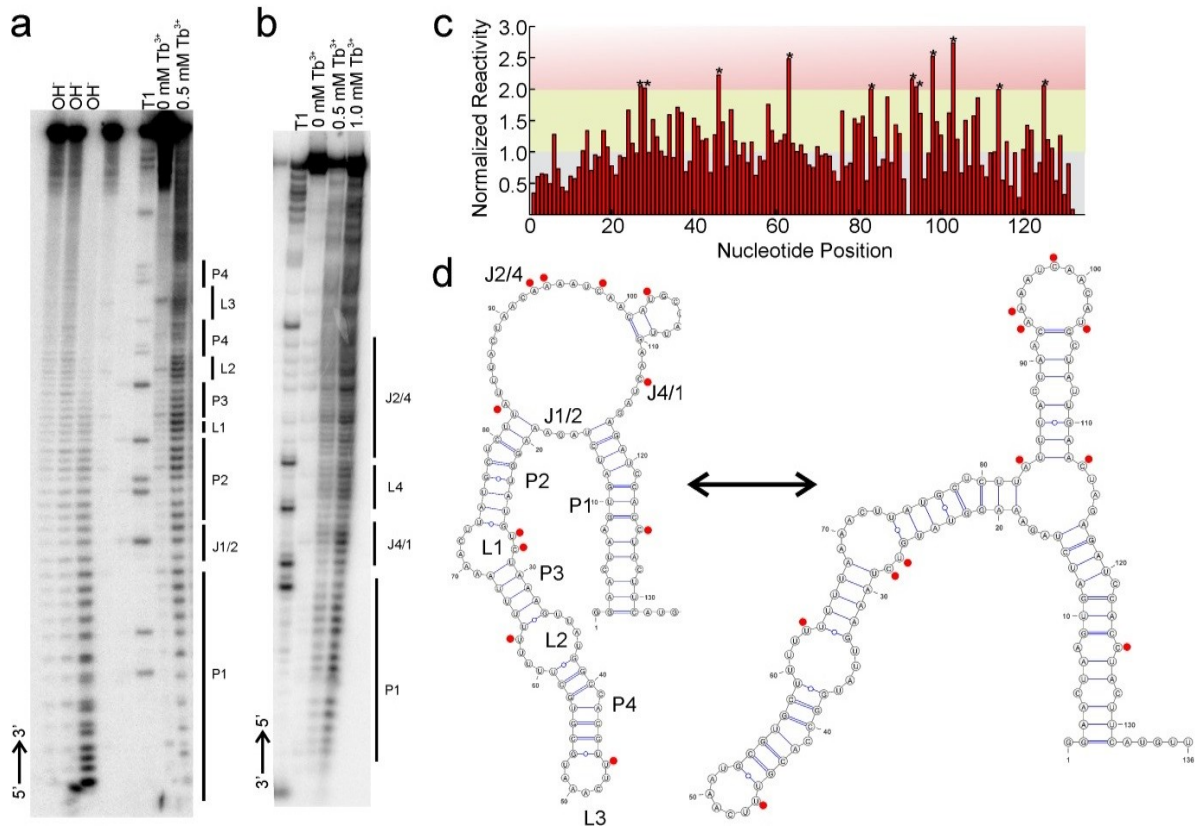



**Figure 4.8 Terbium(III) footprinting of full length Ubc4**
$Tb^{3+}$ footprinting analysis of full length Ubc4 labeled with $^{32}P$ either on the 5' end (a) or $^{32}p$-pCp on the 3' end (b) in the presence or absence of 1.0 mM terbium(III) for 1 hour at 30 °C. $OH^-$: alkaline hydrolysis ladder. T1: G-specific RNase $T_1$. (c) Quantification and normalization of the extent of cleavage for each nucleotide position from at least 3 separate experiments. Values above 2 are considered significant (*) and used as constraints in structure predictions. (d) Footprinting-modified structures of Ubc4 as determined by mFOLD. Reactivity values above 2 used as constraints in structure prediction are marked with a red circle.

demanding approaches to obtain more accurate models of the underlying RNA 3D structures. Accurate prediction of 3D structures is currently limited to smaller RNA molecules primarily due to the fact that biochemical data supporting the output structures is typically lacking. In addition, typical biochemical structure probing, such as footprinting, only reports the average structure of a population of RNA and as a result can mislead in structure prediction.

As steps towards fully developing FRETtranslator, we first implemented the technique for the analysis of a short, 76-nucleotide long RNA containing donor and acceptor fluorophores for smFRET. By individually fitting each smFRET trajectory to idealized FRET states and inputting the resulting time-series FRET paths into FRETtranslator, we were able predict secondary structure trajectories that coincided with the collected smFRET data. FRETtranslator was able to identify alternative high FRET structures that explained the high FRET nature of the substrate. While FRETtranslator predicted the same local substructure for both high FRET states near the nucleotides where the fluorophores were attached, i.e., structure 6438 and structure 1732, other parts of the structures differ in the prediction. The parts that differ do not modulate the donor and acceptor distances. Therefore, smFRET does not yield explicit information about this part of the structures. Nonetheless, we can assume that the 76-nucleotide truncated Ubc4 substrate folds into an elongated shape similar to the predicted mFOLD MFE structure. Similarly, a near-zero FRET state does not provide any distance information. Hence, FRETtranslator could overestimate unfolding of an RNA into a structure like 7401. Still, FRETtranslator predicts that structure 2779 is more likely to show a transition into a low FRET state structure than expected for the high FRET structures discussed above, effectively using our kinetic information from smFRET to inform the predicted RNA folding landscape.

We next applied it to smFRET data of a 135-nucleotide pre-mRNA substrate, Ubc4, alternatively fluorophores labeled in two separate regions of the RNA. In addition to the smFRET information, we have also gathered terbium(III) footprinting data that can be used to further validate and refine the predicted structures. Moreover, the BHG modeling has now been extended to include pseudoknots that will be included for further computations of the full-length Ubc4 substrate. Unfortunately, structure prediction of longer RNA molecules using smFRET data as input for FRETtranslator is still quite time-consuming and has not yet completed the analysis process. However, given enough time, FRETtranslator should prove to be a useful tool for the accurate prediction of potential RNA structures for which smFRET data is known.

We next plan to modify the FRETtranslator algorithm to enable incorporation of footprinting data with which to alter and modify the predicted RNA structures. As most footprinting techniques, such as terbium(III) footprinting, are time- and ensemble-averaged methods and thus can only capture a snapshot of the average RNA structure in solution, the most unreactive and thus most tightly base paired regions of the RNA will be used to modify the structures that are most representative of the most stable FRET conformation. Additionally, the most reactive and thus most single-stranded regions of the RNA will be used to modify these same structures that are most representative of the most stable FRET conformation. For example, the P1 stem of full length Ubc4 is significantly less reactive than the remainder of the structure and thus can be confidently assumed to be base paired. Considering Ubc4-2 primarily adopts a high FRET conformation, FRETtranslator will be programmed to favor high FRET conformation structures in which the P1 stem is formed. In more complex cases, we will develop computational algorithms that time- and ensemble-average the smFRET-guided structures derived from FRETtranslator to calculate partition functions for the base pairing probability of

each nucleotide, which then will be re-calibrated and optimized for convergence with the footprinting data on that nucleotide.

Once the FRETtranslator algorithm has been fully optimized using the sample datasets, we next plan to use previously analyzed smFRET data describing the dynamics of 5'SS and branchsite docking throughout spliceosome assembly up to the first step of splicing[62]. In these experiments, smFRET data were collected using the Ubc4-1 construct in the presence of several mutant yeast splicing extracts that result in the accumulation of particular splicing complexes (**Figure 4.9a**). Interestingly, the Ubc4 substrate was found to undergo very large structural rearrangements, transitioning from a folded, high FRET conformation in CC2, to a near zero, low FRET conformation in the A complex, and then finally to the 0.2 FRET state in the $B^{act}$ complex. Beyond this, little is known about the exact conformation of the pre-mRNA within the spliceosome. To refine our understanding of the spliceosome core and how it changes throughout spliceosome assembly, we plan to use known snRNA and protein binding sites (**Figure 4.9b**)[30,149-152] and the gathered smFRET data for each splicing complex[62] to predict potential 2D and 3D structures of Ubc4 at each stage of assembly. Since simply including the known protection regions for secondary structure prediction by mFOLD produces only one stable structure (**Figure 4.9c**), we expect that the ability to also incorporate smFRET information to refine our predictions will greatly enhance the accuracy of predicting the dynamic time sequence of conformations of Ubc4 as it is processed by the spliceosome.

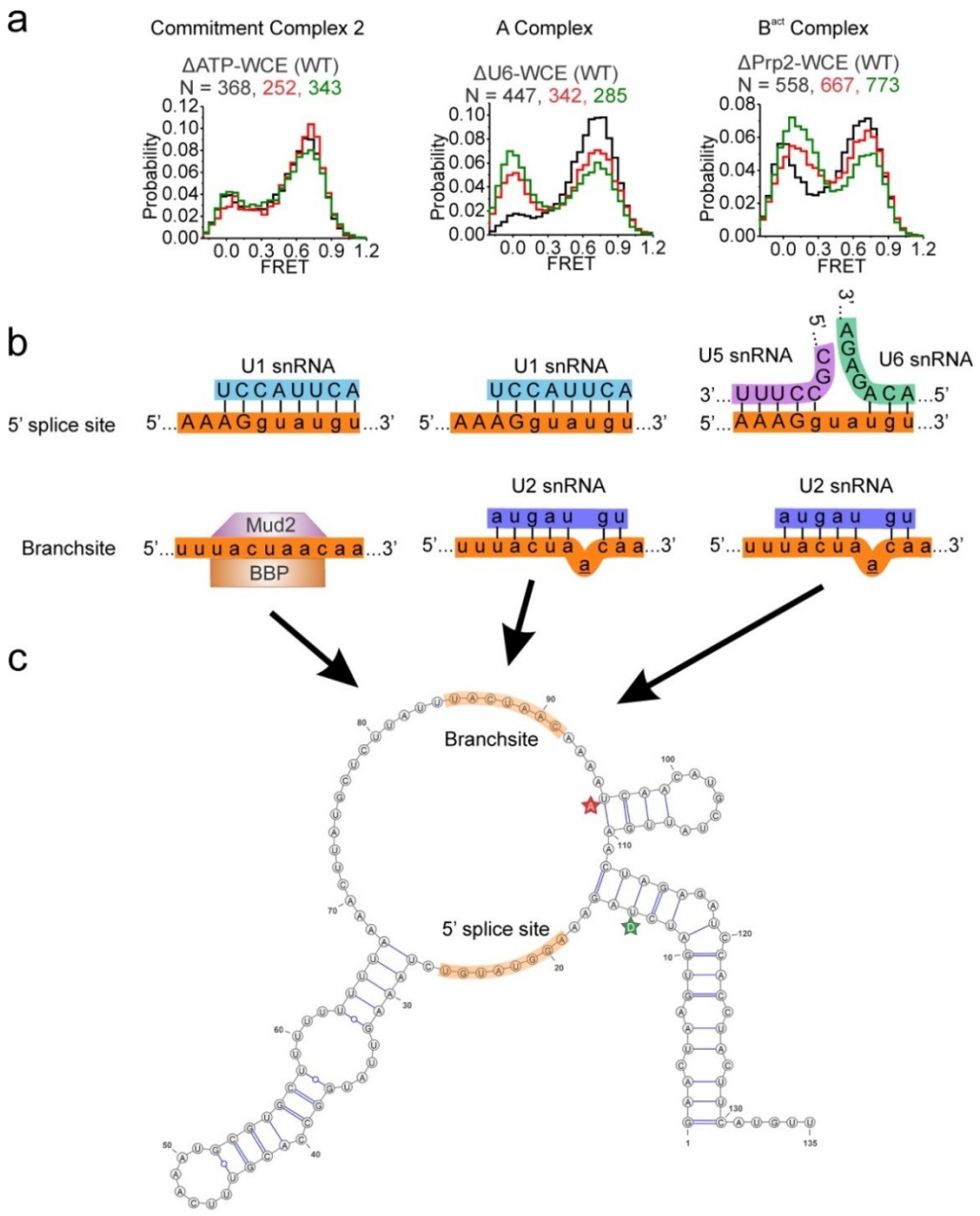



**Figure 4.9 Incorporating known 5'SS and BS protection patterns does not explain observed FRET behavior in several splicing conditions**

(a) Histogram analysis of smFRET data from several splicing conditions known to stall at either commitment complex 2 (CC2), A complex, or the $B^{act}$ complex. A time series was performed in which FRET data was recorded 0-8 min (black), 18-23 min (red), and 33-40 min (green) after addition of the indicated extract condition. (b) Known snRNA and protein binding sites on the 5' splice site and branchsite for CC2, A complex, and $B^{act}$ complex. The pre-mRNA intron (orange) is lowercase while the exon is uppercase (c) mFOLD predicted structure of Ubc4 using the known protection pattern for each complex. Donor (green) and acceptor (red) dyes used in (a) are indicated as well as the 5'SS and branchsite regions.

# CHAPTER 5: Identifying Novel Yeast Introns and Common Secondary Structure Features in an *in vivo* Assembled, Activated Spliceosome[4]

## 5.1 Introduction

The spliceosome is a large macromolecular machine responsible for removing non-coding segments of pre-mRNA (introns) and ligating the flanking coding sequences (exons) to produce the mature mRNA utilized by the ribosome as a template for protein synthesis. This crucial step in the maturation of mRNA allows for the cell- and tissue-specific expression of protein isoforms from a single gene sequence in higher eukaryotes such as humans. To this end, spliceosomes need to reliably identify intron-exon boundaries, excise introns with single-nucleotide precision, discard substrates carrying mutations, and accurately regulate alternative splicing events. Not surprisingly, up to 50% of all mutations leading to human disease are thought to be the result of splicing defects[153].

In contrast to other macromolecular machines such as the ribosome, the spliceosome does not have a pre-formed catalytic core. Rather, each of the five small ribonucleoprotein complexes (snRNPs, denoted U1, U2, U4, U5, and U6), which are themselves composed of a single small nuclear RNA (snRNA) and several associated proteins, assemble on a single pre-mRNA

[4] Matt Kahlscheuer performed biochemical isolation and validation of the *in vivo*-assembled $B^{act}$ complex as well as data analysis confirming the presence of known introns and investigating the presence of new splicing substrates. Nguyen N. (Josh) Vo performed differential expression analysis and assisted with the initial read mapping. Brian Magnuson performed the Tophat and Bowtie analysis to map reads to the yeast genome. Michelle Paulsen performed cDNA synthesis and preparation. Sequencing of the cDNA library was performed by the staff at the University of Michigan Sequencing Core.

substrate in a stepwise fashion, carry out both steps of splicing, and then disassemble to carry out further splicing cycles on other pre-mRNAs (**Figure 5.1b**). Such a stepwise assembly process allows for tight regulation of the splicing process by providing multiple checkpoints before, during, and after both steps of splicing. The splicing cycle begins with the recognition of the 5' splice site (5'SS) and branchpoint (BP) region by the U1 and U2 snRNP complexes, respectively. The U4, U5, U6 snRNPs bind to this A complex structure as a preformed complex known as the tri-snRNP to form the B complex[59]. Large RNA-RNA and RNA-protein rearrangements occur at this point such that the RNA-RNA base pairing between U1 and the 5'SS is disrupted, resulting in the release of U1 from the spliceosome and binding of the U6 snRNP to the 5'SS. These and other RNA and protein rearrangements throughout the cycle are catalyzed by at least eight RNA-dependent ATPases of the DExD/H-box subfamily and are thought to enhance the fidelity of splicing by acting through a kinetic proofreading mechanism such that mutant substrates can be discarded at multiple steps in the assembly pathway[55,65]. Once the U6 and U5 snRNP complexes are stably associated with the spliceosome, the U4 snRNP, which serves more of a cofactor role in preventing premature binding of U6 to the pre-mRNA, dissociates from the spliceosome. This activated form of the spliceosome ($B^{act}$) contains the fully formed catalytic core that, with the help of several of the RNA-dependent ATPases, can carry out the two chemical steps of splicing (**Figure 5.1a**). In the first step, the BP adenosine executes a nucleophilic attack on the 5'SS, resulting in the release of the 5' exon from the newly formed intron lariat structure. Following a further ATP-dependent rearrangement, the 3' hydroxyl of the 5' exon nucleophilically attacks the 3'SS, resulting in formation of the mature mRNA and release of the intron lariat (**Figure 5.1a**).

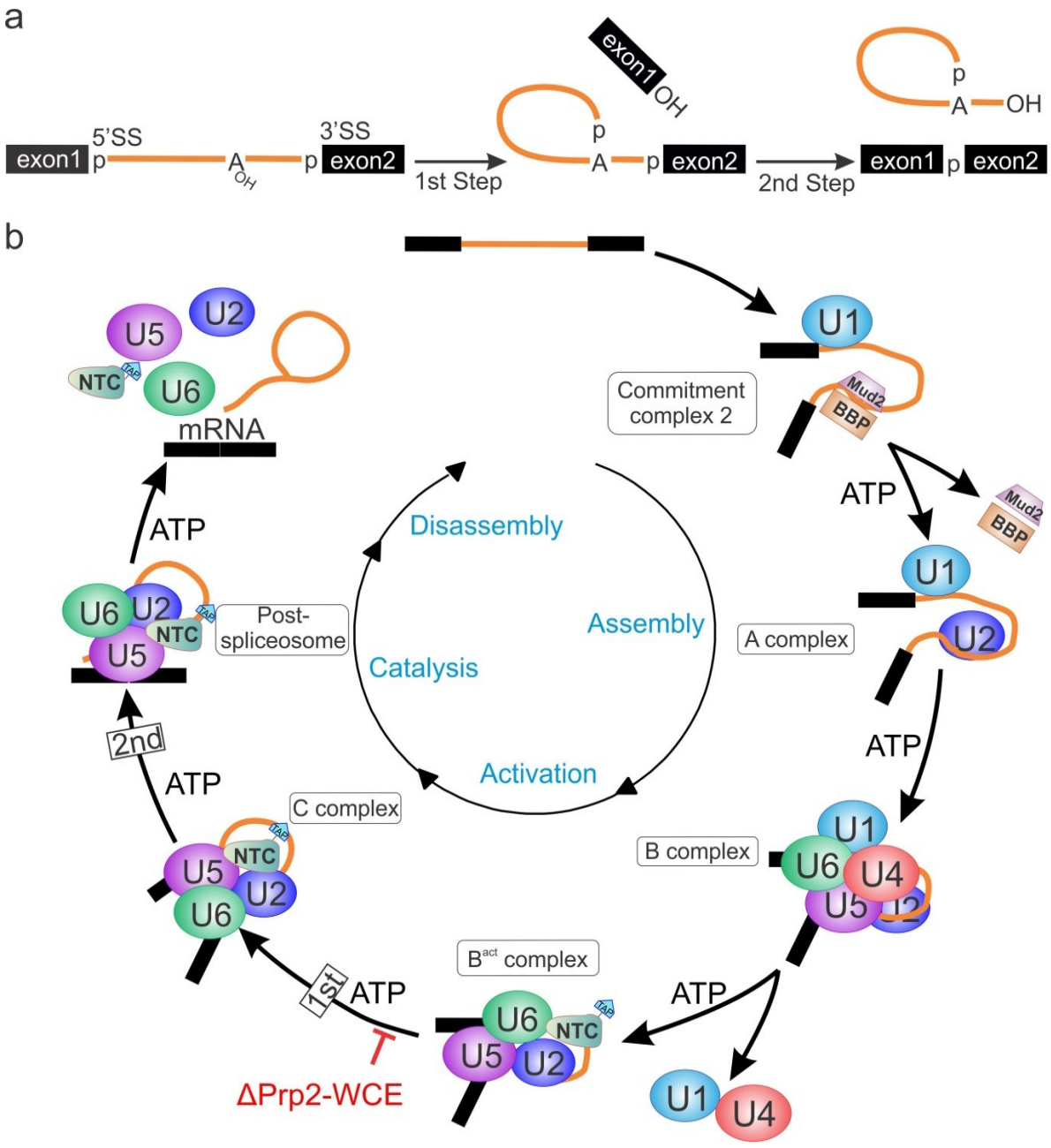



**Figure 5.1 The yeast splicing mechanism and spliceosome assembly pathway**
(a) The two chemical steps of splicing (b) The spliceosome assembly pathway highlighting the Prp2-inactivation resulting in accumulation of the $B^{act}$ complex. The $B^{act}$ complex can be further purified utilizing a TAP-tagged NTC component, Cef1, that associates only at the $B^{act}$ stage and remains throughout the rest of the splicing cycle.

Previous work has shown that, upon incubation with extract, the efficiently spliced Ubc4 pre-mRNA traverses through a unique set of conformational dynamics as it progresses thorough the various stages of splicing. Only upon formation of the activated spliceosomal $B^{act}$ complex do the BP and 5'SS of Ubc4 become stably positioned distal to one another[55]. Initially more surprisingly, in the absence of extract the 5'SS and BP regions are already found in close proximity to one another. This observation indicates that the intron secondary structure is such that these two points of first-step chemistry are brought to within splicing distance before any protein or RNA component of the spliceosome acts on the pre-mRNA, an idea that was first observed in yeast more than 20 years ago[154]. This idea initially is surprising given the large sequence distance between the 5'SS and BP and supports a model in which the intron plays a more active role in positioning the 5'SS and BP close to one another, ready for the first step of splicing, similar to the function of self-splicing group I and II introns[155]. Such a model is, however, in accord with recent studies that appear to support the hypothesis that intron secondary structure has a functional role in splicing. For example, recent *in vitro* experimentation has shown that the RNA secondary structure can influence 5'SS recognition by shortening the 5'SS-BS distance[156]. In addition, others have found that pre-mRNA secondary structure can maintain the 3'SS at the right distance from the BS and modulate the accessibility of the 3'SS to the spliceosome[157]. Lastly, only certain pre-mRNAs such as Ubc4 are efficient spliceosomal substrates *in vitro*, perhaps due to their inherent secondary structure[53].

There is, however, a great deal of resistance to this hypothesis, primarily due to the lack of quantitative studies that correlate RNA secondary structure predictions with experimental conformational dynamics and splicing activity of a specific pre-mRNA. In an effort to delineate the functional impact of intron secondary structure on splicing activity, we have isolated the *in*

*vivo* assembled, activated yeast spliceosome containing nearly all known yeast pre-mRNA substrates. Deep sequencing analysis of the RNA within the complex reveals >90% of the known pre-mRNA substrates to be present in significant abundance. Furthermore, we have identified a number of previously unknown pre-mRNA substrates that may be subject to cellular regulation by the spliceosome-mediated decay (SMD) pathway. With further investigation using the newly developed SHAPE-MaP technology, we plan to correlate common secondary structure features with splicing efficiency to rigorously test our hypothesis of a relationship between extent of pre-mRNA secondary structure and inherent splicing activity. If common structural features are detected in efficiently spliced substrates, these could serve as future targets for small-molecule drugs to influence splicing outcomes as well as the downstream translation/degradation/transport of the resulting mRNAs[158].

## 5.2 Materials and Methods

### 5.2.1 $B^{act}$ complex enrichment and purification

The *in vivo*-assembled $B^{act}$ complex was isolated from a *prp2-1 cef1-TAP* yeast strain (ATCC 201388: *MAT***a** *his3*Δ*1 leu2*Δ*0 met15*Δ*0 ura3*Δ*0*[71]), heated at 37 °C for 40 min to inactivate Prp2 and stall the spliceosome at the $B^{act}$ complex essentially as described[39]. In brief, the mutant strain was grown at the permissive temperature of 25 °C to an OD ($A_{600}$) of 1.0 and then shifted to the non-permissive temperature of 37 °C for 45 min to inactivate Prp2 and allow buildup of the $B^{act}$ complex. Cells were harvested by centrifugation, washed with buffer A (10 mM HEPES-KOH pH 7.9, 200 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM DTT, 10% (v/v) glycerol, 0.6 mM PMSF, and 1.5 mM benzamidine), and resuspended in one volume (w:v) buffer A before freezing dropwise in liquid nitrogen to form small cell pellets that could be stored at -80 °C.

Frozen cells were then milled to make splicing active whole cell extract (WCE) as described[53,121]. Briefly, cell pellets were disrupted by manual grinding using a mortar and pestle half-submerged in liquid nitrogen for 20-30 min before thawing in an ice bath. ATP was depleted by addition of glucose to 2 mM. Insoluble material was pelleted at 17,000 rpm in a type 45 Ti Beckman rotor followed by a second centrifugation step at 37,000 rpm in a Ti-70 rotor for 1 h. The clear middle layer was removed with a syringe and dialyzed for 4 h against 20 mM HEPES-KOH, pH 7.9, 0.2 mM EDTA, 0.5 mM EDTA, 150 mM KCl, 20% (v/v) glycerol, 0.1 mM PMSF, and 0.25 mM benzamidine with one buffer exchange. The resulting extract (~40 mL) was incubated with IgG-sepharose (500 µL) with rotation at 4 °C for 2 h. The resin was washed thoroughly with ~25 mL of wash buffer (10 mM Tris-HCl pH 8.0, either 150 mM (low salt sample) or 500 mM (high salt sample) NaCl, 0.1% NP-40, 1.5 mM $MgCl_2$, 8% (v/v) glycerol, 1 mM DTT, 0.2 mM PMSF). The resin-bound $B^{act}$ complex was then eluted through incubation with a 500 µL reaction containing TEV protease and RNase inhibitor for 3 h at 16 °C. Eluted material was layered onto a 10%-30% glycerol gradient containing 20 mM HEPES-KOH, pH 8.0, 150 mM KCl, 1.5 mM $MgCl_2$, and 0.1% NP-40 and centrifuged for 10-14 h at 29,000 RPM in an SW41 rotor. Fractions (450 µL) were collected from the top with a pipette, phenol/chloroform extract to remove protein, and ethanol precipitated to isolate the bound RNA. RNA was analyzed by Northern blot analysis probing for U1, U2, U4, U5, and U6 snRNAs, RT-PCR of the Act1 pre-mRNA, and deep sequencing analysis.

5.2.2 Northern blot and RT-PCR analysis

Precipitated RNA from the even glycerol gradient fractions was resolved on a 7M urea, 6% (v/v) polyacrylamide gel and transferred to BrightStar®-Plus Positively Charged Nylon Membrane (Life Technologies) for 60 min at 200 mA. Membranes were dried and pre-hybridized for 30 min

at 37 °C in RNA hybridization buffer (50% (v/v) formamide, 0.5% (w/v) SDS, 5X Denhardt's solution (Life Technologies), 10 μg/μL Salmon sperm DNA, 750 mM NaCl, 50 mM $NaH_2PO_4$, 5 mM EDTA). snRNA probes (**Table 5.1**) were 5' end labeled with $^{32}P$ using T4 Polynucleotide Kinase (NEB) and purified using Centri-spin columns (Princeton Separations). Labeled probes were added to membrane and rotated at 42 °C overnight. Membranes were washed with 3-4 times at 37 C for 10-20 min each with 10 mL of wash buffer (30 mM sodium citrate, 300 mM NaCl, and 0.1% (w/v) SDS) before exposing to a phosphor screen and scanning on a Typhoon variable mode imager (GE Healthcare).

RT-PCR first strand synthesis was performed by mixing 2.5 μM of the RT primer with up to 5 μg of RNA, denaturing at 70 °C for 5 min, and cooling in an ice bath. The primer-RNA mixture was combined with Transcriptor reverse transcriptase (Roche Life Science) according to the manufacturer's protocol.

5.2.3 cDNA library preparation and Illumina Hi-Seq sequencing

Isolated RNA contained within the $B^{act}$ complex was converted into a strand-specific DNA library using the Illumina TruSeq Kit and size selected at around 200 base pairs as previously described[159,160]. Briefly, RNA was fragmented at 85 °C for 10 min prior to first strand cDNA synthesis in the presence of Actinomycin D to yield strand specific reads which were then purified using AMPure RNAclean beads (Beckman Coulter). The second strand cDNA was then synthesized and the resulting cDNA was purified with AMPure XP beads. An Illumina TruSeq RNA Sample Prep Kit was used to repair the purified cDNA ends, adenylate and ligate adaptors to the cDNA. The samples were then gel purified on a 3% agarose gel and size-selected by excising the 200 bp region of the gel and isolating the cDNA using the QIAEX II Gel Extraction

| | Sequence |
|---|---|
| Anti-U1 | CACGCCTTCCGCGCCGT |
| Anti-U2 | CTACACTTGATCTAAGCCAAAAGGC |
| Anti-U4 | AGGTATTCCAAAAATTCCC |
| Anti-U5 | AAGTTCCAAAAAATATGGCAAGC |
| Anti-U6 | ATCTCTGTATTGTTTCAAATTGACCAA |
| YFL039C_m2_F | TCGAAAATTTACTGAATTAACA**ATGGA** |
| YFL039C_E4_R | GATGGGAAGACAGCACGAGGAG |
| YFL039C_L_R | *GCAAGCGCTAGAACATAC* **A**TAGTACA |

**Table 5.1 snRNA and RT-PCR DNA oligonucleotides used in the study**
Primer m2_F binds opposite the Act1 5'UTR and the first five protein coding nucleotides (bold). Primers E4_R and L_R are used for Act1 cDNA synthesis using Reverse Transcriptase. Primer L_R is used to amplify first step splicing products: bold sequence is opposite the branchpoint adenosine; underlined is complementary to the region upstream of the branchpoint; italic sequence is complementary to the 5'SS.

Kit (Qiagen). Finally, TruSeq Kit PCR reagents were used to enrich the DNA fragments before a final purification steup using AMPure XP beads.

Sequencing of the cDNA library was performed by the staff at the University of Michigan Sequencing Core using the Illumina HiSeq 2000 sequencer. Base calling was performed using Illumina Casava v1.8.2. and read mapping was performed by first mapping to ribosomal RNA using Bowtie analysis followed by mapping of the remaining reads to sacCer3 reference genome using Tophat. Expression levels for exons, introns, and whole genes were calculated using the RPKM unit: $RPKM = \frac{(10^9 \times C)}{N \times L}$, where C = Number of reads mapped to a gene, N = Total mapped reads in the experiment, and L = exon length in base-pairs for a gene. Data of mapped reads were plotted using a custom-built browser as previously described[160].

5.2.4 Differential expression analysis

Differential analysis was performed using Cuffdiff essentially as described[161]. Cuffdiff is a statistical analysis tool that takes aligned reads from two or more experimental conditions and determines genes and transcripts that are differentially expressed using a linear statistical model. Cuffdiff uses several biological replicates from each condition being analyzed to learn how expression levels of genes vary across replicates and calculates a significance of observed changes using these variance estimates.

Following TopHat-mediated mapping of the RNA-seq reads, Cuffdiff was run using R Programming. Results were reported as a set of tab-delimited text files which were used for further analysis and plotting using Microsoft Excel.

## 5.3 Results

### 5.3.1 Isolation of the *in vivo* assembled $B^{act}$ complex

In recently published work, a method was described which utilized two yeast genetic modifications for the purification of the yeast $B^{act}$ complex containing a FRET labeled substrate Ubc4, a technique termed Single Molecule Pulldown FRET (SiMPull-FRET)[55]. The first modification is a heat-sensitive mutation in the ATPase Prp2 known to cause inactivation of the protein upon incubation at the non-permissive temperature[37]. Prp2 is a member of the DEAD/H box family of RNA-dependent ATPase known to function in an ATP-dependent manner immediately prior to the first step of splicing. As a result of its inactivation, spliceosome assembly stalls at the activated $B^{act}$ complex (**Figure 5.1b**). The second modification is the insertion of a Tandem Affinity Purification tag (TAP-tag) into the protein Cef1. Cef1 is an essential splicing factor associated with the Nineteen complex that is known to associate with the spliceosome only upon formation of the $B^{act}$ complex and remain bound through both steps of splicing[39]. Utilizing these mutations the authors were thus able to stall spliceosome assembly specifically at the $B^{act}$ stage and then isolate this complex from the remaining immature splicing complexes by incubating the extract with magnetic beads coated with IgG, which specifically binds the TAP-tag.

To adapt this purification method to isolating the $B^{act}$ complex containing all yeast pre-mRNA substrates, we made a number of modifications similar to the previously described biochemical approach[39] (**Figure 5.2a**). Specifically, the Prp2 and Cef1 modified yeast strain was grown at the permissive temperature of 25 $^{\circ}$C and then shifted to the non-permissive temperature of 37 $^{\circ}$C for 40 min prior to harvesting to allow for the inactivation of Prp2 and trapping of endogenous pre-mRNAs in a first-step arrested spliceosome. RT-PCR analysis of the ACT1

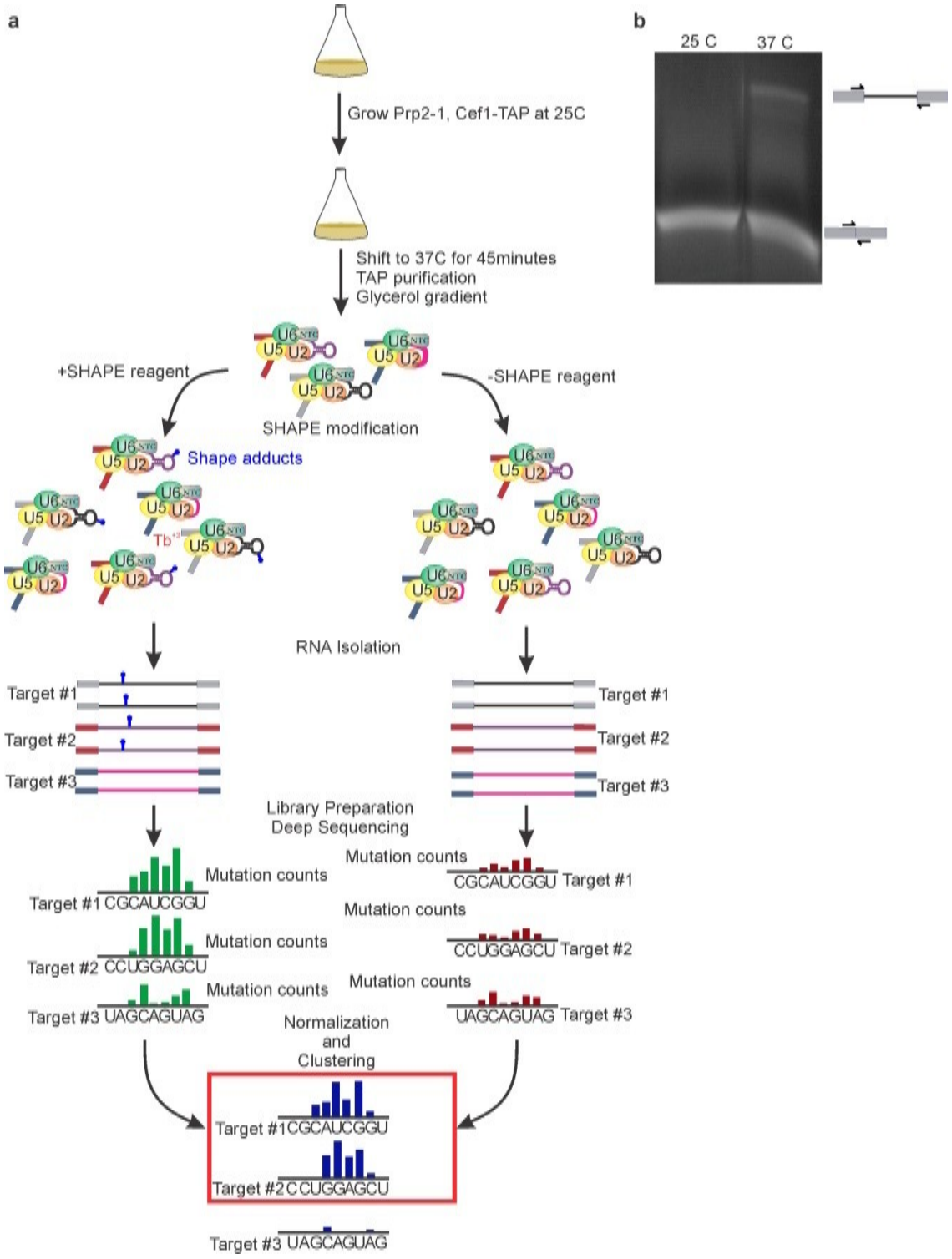



**Figure 5.2 Workflow of $B^{act}$ isolation and SHAPE-MaP profiling**
(a) Workflow for the isolation of the *in vivo* assembled $B^{act}$ complex and subsequent SHAPE-MaP analysis (b) RT-PCR of Actin RNA confirming presence of primarily mRNA prior to heating of extract and the accumulation of pre-mRNA after shifting to 37 °C.

RNA, a highly abundant pre-mRNA substrate for the spliceosome, within the pre- and post-inactivated culture revealed a small amount of accumulation of ACT1 pre-mRNA when heated to 37 °C (**Figure 5.2b**) indicating successful inactivation of splicing. Whole cell extract (WCE) was prepared from this growth, incubated with IgG-sepharose beads to allow binding of stalled spliceosomes, and the resin washed thoroughly with either a low salt (150 mM NaCl, LS) or high salt (500 mM NaCl, HS) buffer to remove unincorporated splicing factors and pre-mRNA. Bound complexes were removed using TEV protease and further purified by layering onto glycerol gradients as described[39] (**Figure 5.2a**). RNA from the even gradient fractions was isolated and checked by Northern blot analysis for the presence of spliceosomal snRNAs known to be present in the activated spliceosome. A significant peak of U2, U5, and U6 snRNAs was detected near ~40S (**Figure 5.3a**), corresponding to the size of the yeast spliceosome[64]. In addition, RT-PCR analysis of the ACT1 RNA within the complex revealed elevated levels of pre-mRNA, indicating successful isolation of a fully assembled, pre-first step spliceosome (**Figure 5.3b**).

5.3.2 The $B^{act}$ complex contains nearly all known pre-mRNA substrates

Recent microarray analysis has shown that more than 80% of the over 250 intron-containing pre-mRNAs in yeast show elevated levels upon inactivation of Prp2[162]. Additionally, a number of studies utilizing RNA-seq have attempted to confirm or refute the presence of several predicted introns and also discover new intron-containing genes or genes associated with the spliceosome[163,164]. The first of the latter studies performed cDNA analysis (RNA-seq) of full-length, 5'-capped mRNA in an effort to completely annotate the yeast transcriptome[163]. Sequencing of two cDNA libraries allowed for the identification of new transcription start sites (TSSs), open reading frames (ORFs), and 45 previously undescribed introns, including several

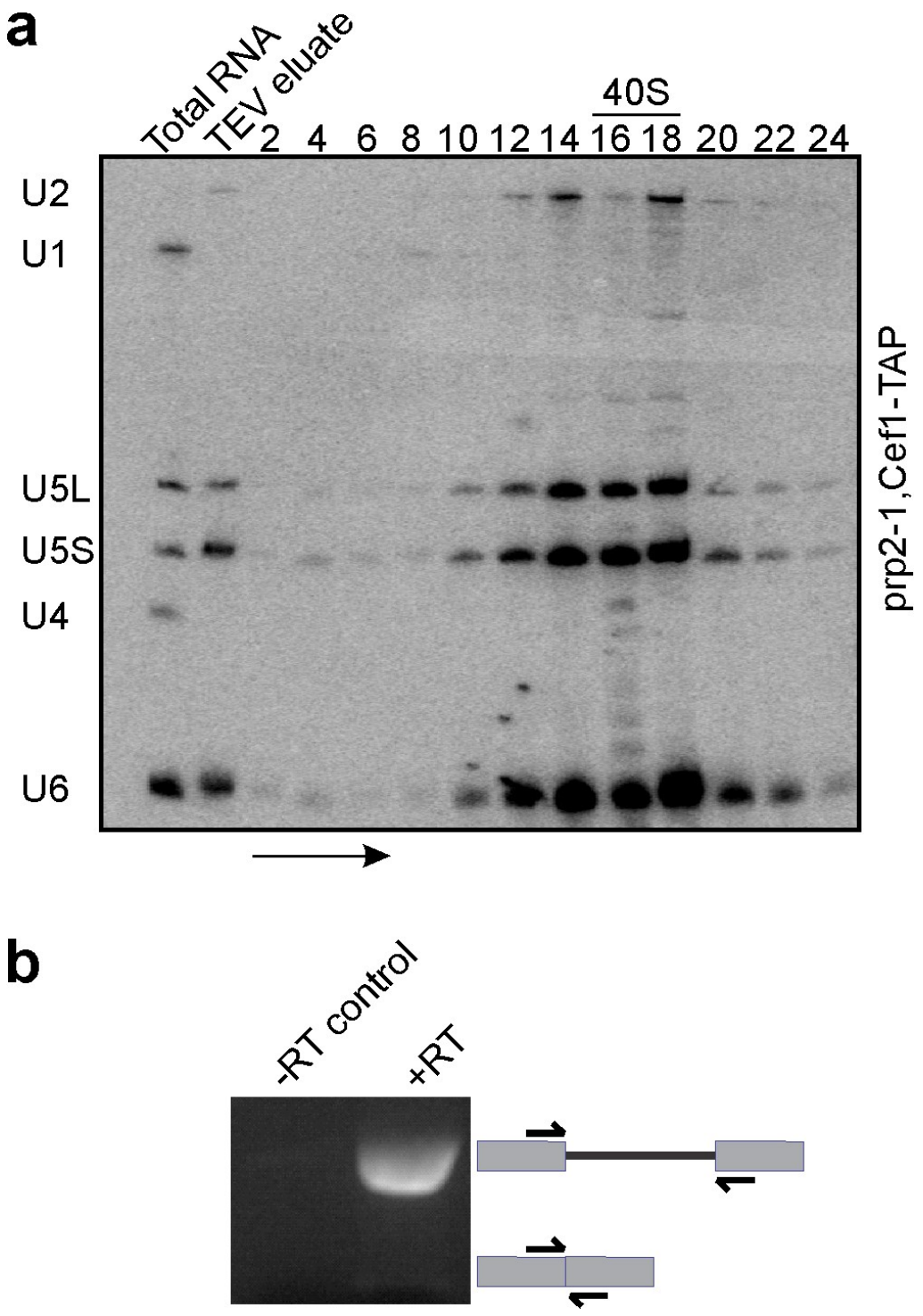



**Figure 5.3 Purification of Prp2-arrested spliceosomes**
(a) Cef1-TAP-associated material was Northern blotted for spliceosomal RNAs. Splicing complexes sediment ~40S as indicated by the presence of U2, U5, and U6 snRNAs. (b) RT-PCR analysis of RNA within the 40S fraction using primers designed to amplify ACT1 RNA. RT = reverse transcriptase.

affecting current ORF annotations. Many of the introns discovered have since been incorporated into the Yeast Genome Database[165] as well as the Ares Lab Intron Database[166]. The second approach isolated SmD1-associated RNA from growing yeast and analyzed bound RNA using RNA-seq. SmD1 is a protein required for stabile binding of the U1 snRNA to the pre-mRNA[167] and thus is only expected to pull down the spliceosomal A complex. Such an approach allowed the authors to identify 60% of known spliced mRNAs, potentially showing under-enrichment of intron-containing genes due to rapid splicing of many transcripts. Unfortunately, this method is far too non-specific as nearly 200 non-intronic genes were identified in their complex, many of which potentially contain Sm-binding sites and thus have little association with the spliceosome. The authors did, however, discover a novel pathway by which some non-intronic genes (such as BDF2) are down-regulated in what they termed the Spliceosome Mediated Decay (SMD) pathway.

To first confirm the presence of these known and predicted pre-mRNA substrates[165,166] we prepared and submitted the $B^{act}$-associated RNA for deep sequencing analysis (RNA-seq). We first investigated whether the proposed and confirmed intron-containing genes (ICGs) were of higher abundance in our purified complex compared to that of the other ~6,600 known genes in yeast. Box plots of the RPKM values for either intron containing genes (ICGs) or genes without an intron (non-ICGs) were generated with the non-ICGs RPKM values taken from the low-salt purified $B^{act}$ complex considering it is the only sample for which there is a replicate. As expected, the ICGs in both the high salt purified (HS) and low salt purified (LS) $B^{act}$ complexes are of significantly higher RPKM values compared to that of the non-ICGs and appear to reliably overlap in RPKM values (**Figure 5.4**). This would support the notion that the ICGs within the $B^{act}$ complex are actually tightly bound by the spliceosome and are thus not affected by the

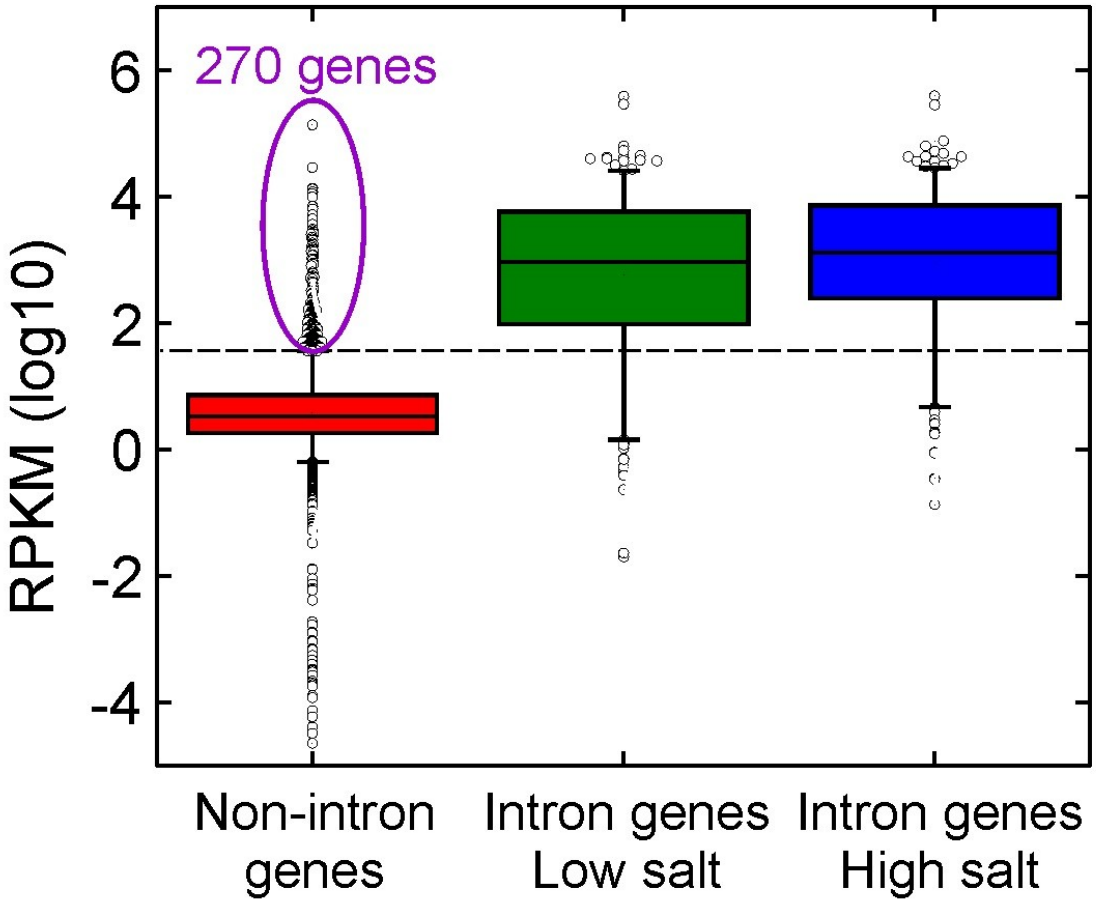



**Figure 5.4 Genes predicted to contain introns have elevated RPKM values relative to non-intron containing genes**

Box plot analysis highlighting the 95%, 75%, mean, 25%, and 5% quartiles for non-intron containing genes and intron containing genes within the purified $B^{act}$ complex purified under low salt (150 mM NaCl) or high salt (500 mM NaCl) conditions. The 270 non-intron genes are those above the 1.6 RPKM (log10) threshold at the 95% quartile.

presence of increased ionic strength during the purification. Using a threshold RPKM value ($log_{10}$) of 1.6 (the 95% quartile for the non-ICGs) encompasses 264 of the 328 (80%) known or predicted intron-containing genes gathered from all known sources[163-166,168] indicating robust enrichment of ICGs in the $B^{act}$ complex. Interestingly, a number of the predicted intron-containing genes are meiotic genes known to be expressed primarily during meiosis and thus are not expected to be within our complex[169]. The majority of these meiotic genes are, in fact, absent from our $B^{act}$ complex, further improving the percentage of intron-containing genes detected to 86% (264/311). Furthermore, a majority of the remaining undetected intron-containing genes are either present in telomeric regions of the chromosome and thus scarcely transcribed, or are mitochondrial RNAs that were not mapped in this study, resulting in a final detection of 264 out of the 283 introns predicted to be in our sample (93%). The remaining intron-containing genes may be expressed at too low of a level in the conditions we are using to be identified in our assay.

Several of the introns detected in the $B^{act}$ complex have not yet been included on the Yeast Genome Database, possibly due to a lack of experimental confirmation. Our data support the inclusion of several predicted introns, including: YKL133C, YMR147W, YNL194C, YGL136C, YER167W, YGL063W, and YMR148W. In addition, the lack of detection of a number of introns, specifically the telomere pre-mRNAs and several introns only recently predicted[163], may be evidence for their withdrawal from the list of genes possessing an intron. However, our detection might be specific to the experimental conditions used and thus further experimentation is required for their removal as intron-containing genes.

5.3.3 Small nucleolar RNAs dissociate from the spliceosome in the presence of high salt

Analysis of the greater than 6,000 genes predicted to be without introns revealed over 270 genes with an RPKM value over 1.6 and thus showing significant detection in the $B^{act}$ complex (**Figure 5.4**). Among this group of genes are, interestingly, 35 of the 73 known small nucleolar RNAs (snoRNAs). snoRNAs are small, stable RNAs found within ribonucleoprotein complexes (snoRNPs) in the nucleoli of eukaryotic cells primarily responsible for the modification of ribosomal RNA (rRNA). Aside from snR17a and snR17b, the snoRNAs are not thought to contain introns and thus are not thought to be recognized as spliceosomal substrates. We thus looked for the presence of these snoRNAs in $B^{act}$ complexes purified in the presence of high salt (HS purified), a condition that should remove any non-specifically bound RNA. Notably, the presence of 500 mM NaCl during purification results in near complete removal of all snoRNAs from the $B^{act}$ complex (**Figure 5.9**), supporting the hypothesis that snoRNA association results primarily from weak interactions with other RNA molecules present in the preparation (such as the ribosomal RNA), not through association with the spliceosome. Removal of the snoRNAs from the statistics results in ~230 non-ICGs present in the $B^{act}$ complex of similar abundance to those of the ICGs (**Figure 5.6**).

5.3.4 Differential analysis identifies a number of novel pre-mRNA substrates

Further investigation of the approximate 230 remaining non-ICGs with high RPKM values for the presence of potential splice sites would be tedious and time-consuming. In addition, the high RPKM values of many of these genes could simply be due to the fact that they are a highly expressed gene and thus there is a significant amount of the RNA present in the yeast cell prior to purification. We therefore employed the use of differential expression analysis to look for genes that are not only of high RPKM values, but that also show elevated levels

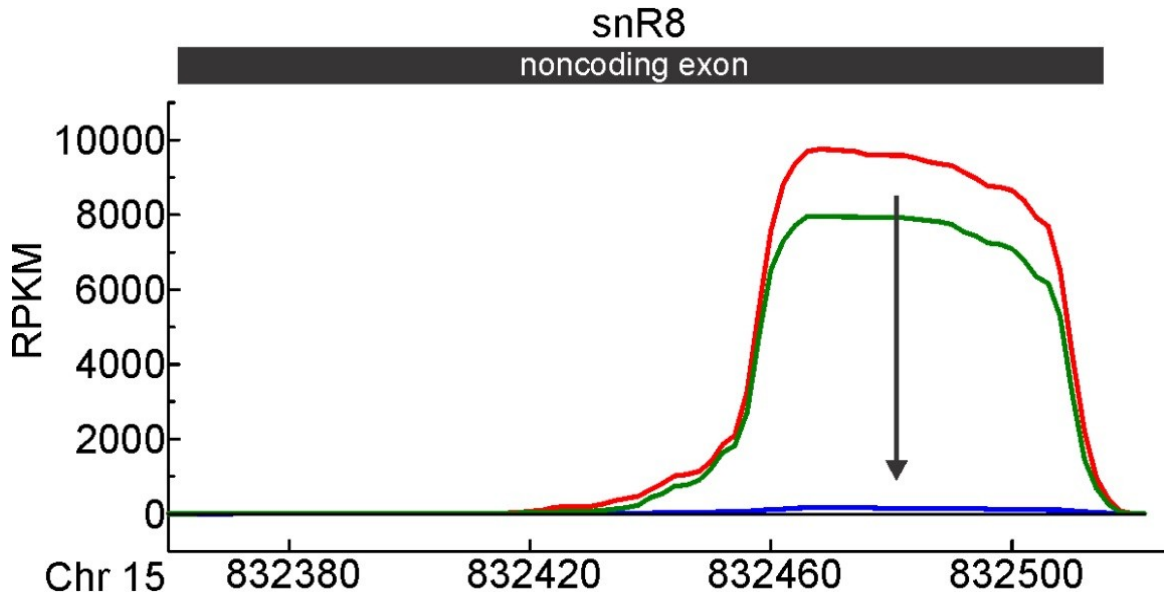



**Figure 5.5 Washing the $B^{act}$ Complex with high-salt buffer results in dissociation of the snoRNAs**

The $B^{act}$ complex was washed with either 150 mM (green and red curves) or 500 mM NaCl (blue curve) during purification. Shown are the mapped reads for snR8 as an example of the removal of snoRNA during high salt purification.

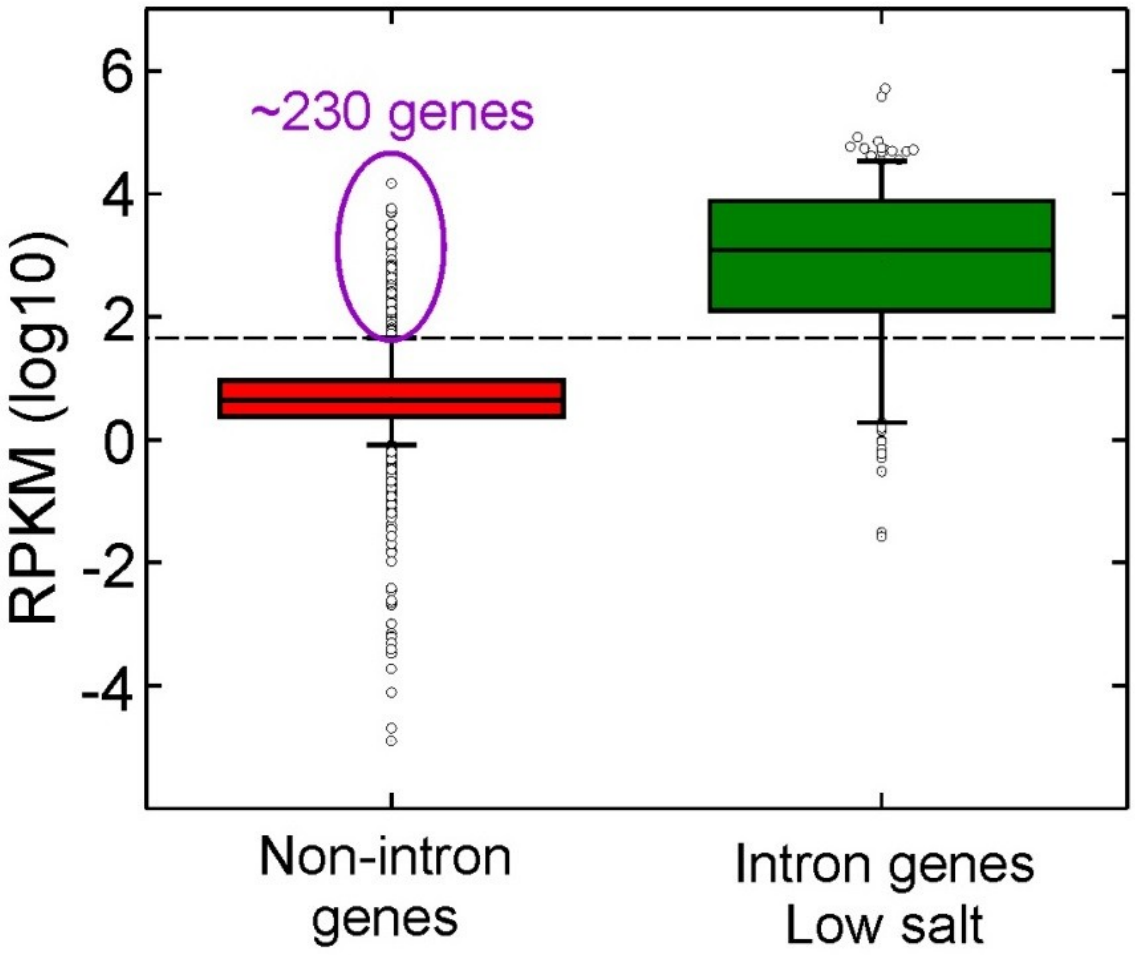



**Figure 5.6 Exclusion of snoRNAs yields ~230 genes not predicted to contain an intron with high RPKM values**

snoRNAs were removed from the list of non-ICGs yielding ~230 non-ICGs above the 1.6 RPKM threshold. The dotted line indicates the 1.6 RPKM threshold at the 95% quartile.

within the $B^{act}$ complex relative to the level of the gene within the cell. Differential expression analysis compares the relative levels of individual RNAs with the corresponding levels in a reference sample[161]. In the case of the $B^{act}$ complex, the reference sample is the RNA from the extract from which the complex was isolated. For such an analysis, an algorithm calculates a fold-change in RNA levels and assigns a significance value (p value) when a number of replicates are used. Unfortunately, we currently only have a replicate for the LS purified $B^{act}$ complex but not for the extract total RNA sample or HS purified $B^{act}$ complex. As a consequence, we are not able to confidently assign significance for any of the genes during this analysis. However, we are able to use the fold-change values as a method to narrow the number of candidate genes containing potential introns.

Fold-change values were calculated for each detected gene in the sample by looking at the ratio of total RNA-to-$B^{act}$ RNA. As a result, genes that show an increased presence in the $B^{act}$ complex sample will result in a very small fraction and thus produce a large, negative value when converted to a $\log_2$ scale as is conventionally done. Again, box plots were developed for either genes not predicted to contain introns or genes predicted to contain introns. As expected, intron-containing genes were found to possess a much larger negative fold change in both the HS and LS purified $B^{act}$ complex relative to the non-intron containing genes (**Figure 5.7**). Using a threshold for the 0.95 percentile of non-intron containing genes ($\log_2$ of -1.49, ~2.81-fold enriched in the $B^{act}$ complex over extract) encompasses 94% (265/283) of the intron-containing genes predicted to be within the $B^{act}$ complex (again excluding mietotic, mitochondrial, and telomeric genes). Unfortunately, this analysis alone still predicts over 240 non-intron containing genes as being enriched in the $B^{act}$ complex (**Figure 5.7**).

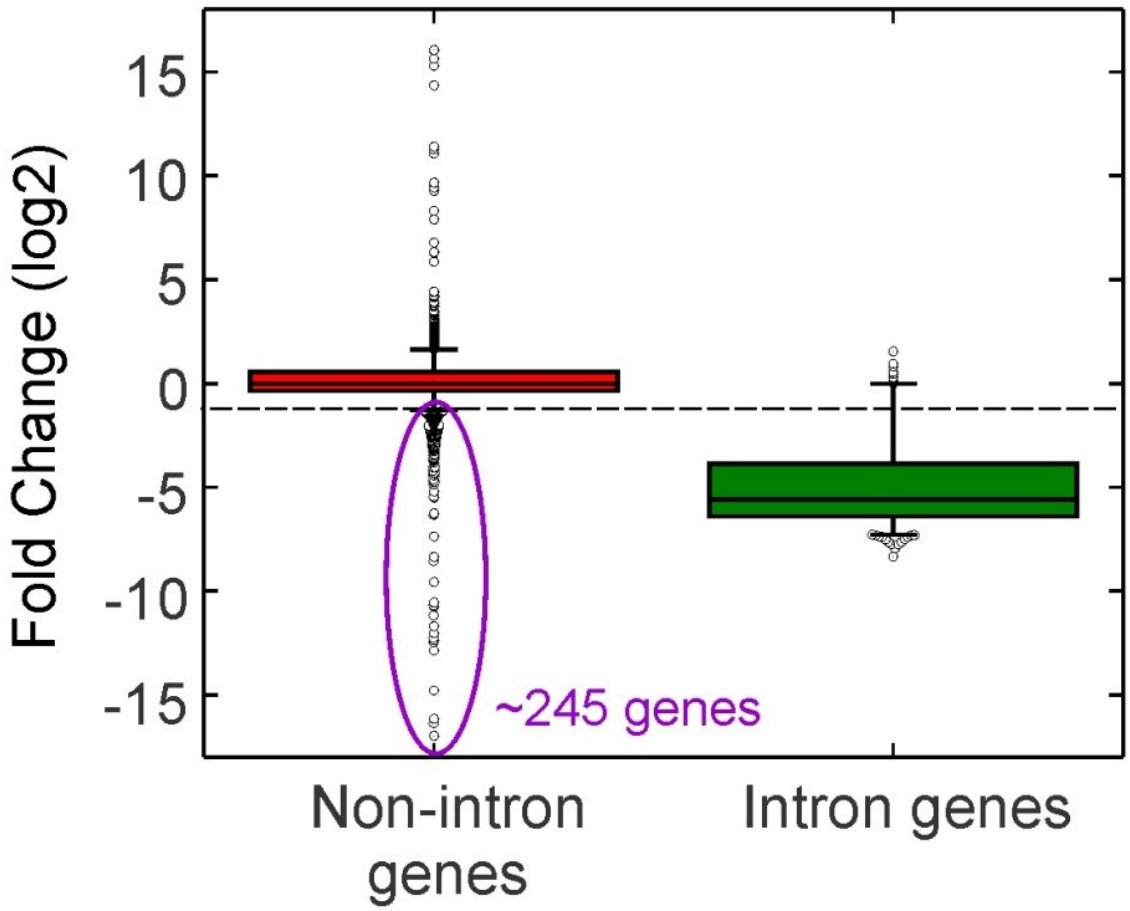



**Figure 5.7 Intron-containing genes show strong enrichment in the $B^{act}$ complex**
Differential expression analysis of the LS purified intron-containing genes and all other genes in the $B^{act}$ complex relative to levels in a total extract RNA sample. snoRNAs were removed from the list of genes not predicted to contain an intron and a threshold of -1.5 (95% quartile, dotted line) was set yielding 245 non-intron coding genes showing significant enrichment in the $B^{act}$ complex.

In an effort to further narrow down our search of new pre-mRNA substrates, we analyzed RNA transcripts that show both a high RPKM value across the gene and a significant increase in abundance (large $-\log_2$ value) in the $B^{act}$ complex relative to yeast extract. Such an analysis shows clear enrichment of the intron-containing genes within the $B^{act}$ complex (**Figure 5.8, green dots).** Utilizing the RPKM and fold change thresholds of 1.6 and -1.49, respectively, still accounts for 265 of the intron-containing genes. In addition, the number of non-ICGs is now reduced to a much more manageable ~70 genes (**Figure 5.8, red dots in upper left quadrant**). Interestingly, a large majority of the resulting non-ICGs overlap on the same strand with all or a significant portion of many of the proposed intron-containing genes (**Figure 5.9a**). It is most probable that these reads were actually incorrectly mapped and belong to the ICG that it overlaps. After removal of the non-ICGs that overlap with a known or predicted intron gene, we were left with 13 previously unidentified intron-containing genes (**Figure 5.9b** and **Table 5.2**). It should be noted that all of these genes, as well as the detected intron-containing genes, have similar or greater RPKM values in the high salt purification indicating tight association with the spliceosome.

Interestingly, three of the $B^{act}$-associated non-ICGs (*YMR148W, YNL195C*, and *YJL206C*) were recently shown to be involved in splicing with an upstream intron-containing gene[163]. *YMR148W* was previously found to have the potential to be expressed as a single transcript with the upstream ORF *YMR147W*, a known ICG. In addition, this work found that *YMR147W* appears to act more as an upstream exon for *YMR148W* and can be spliced to form a hybrid protein between *YMR147W* and *YMR148W* (**Figure 5.10a**). Our data appear to match these results as the two genes are detected in the $B^{act}$ complex as a single transcript with strong peaks near the sites of potential transcription and splicing (**Figure 5.10b**). *YNL195C* encodes a

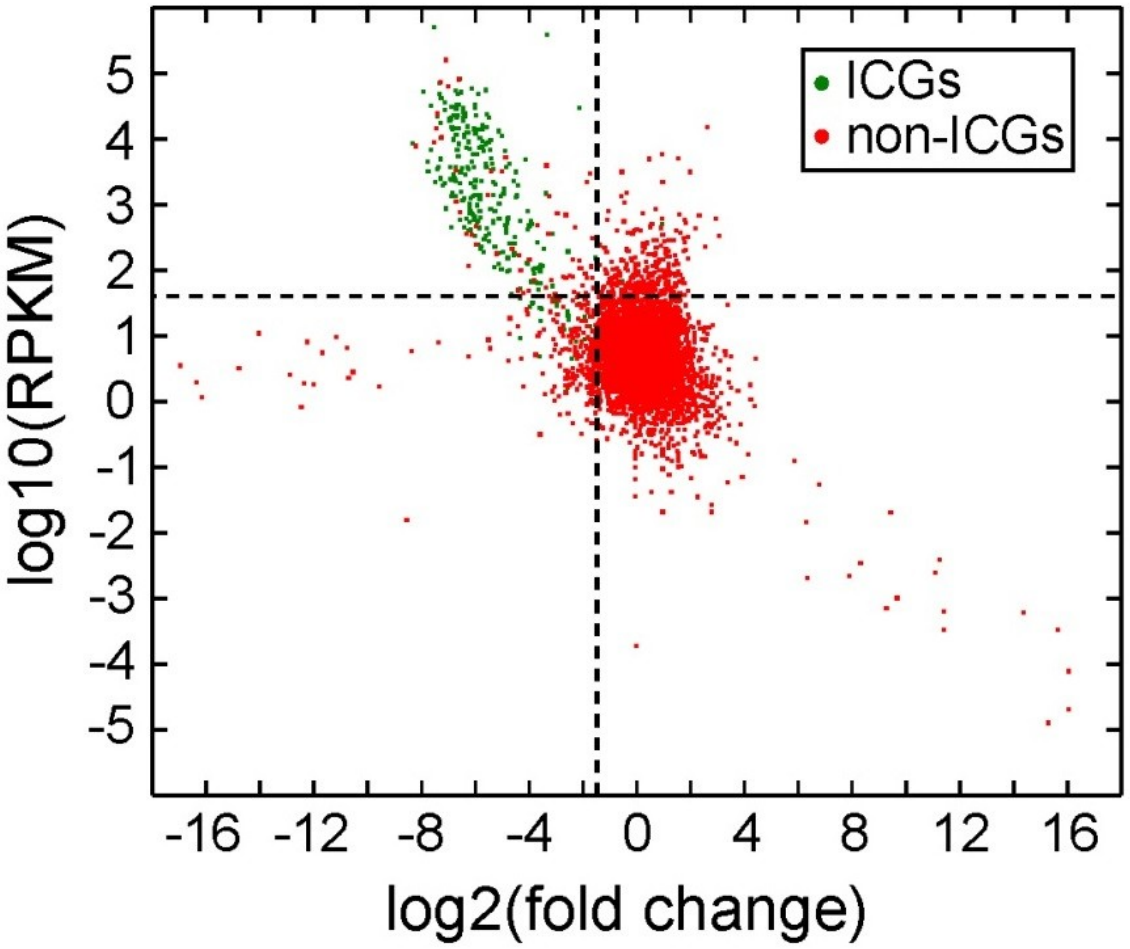



**Figure 5.8 ICGs and several non-ICGs are enriched and have high RPKM in $B^{act}$**
Scatter plot depicting RPKM and fold change relative to total extract RNA values for ICGs and non-ICGs. Vertical and horizontal lines indicate the fold change and RKPM thresholds of -1.5 and 1.6, respectively. snoRNAs have been removed from the possible non-ICGs.

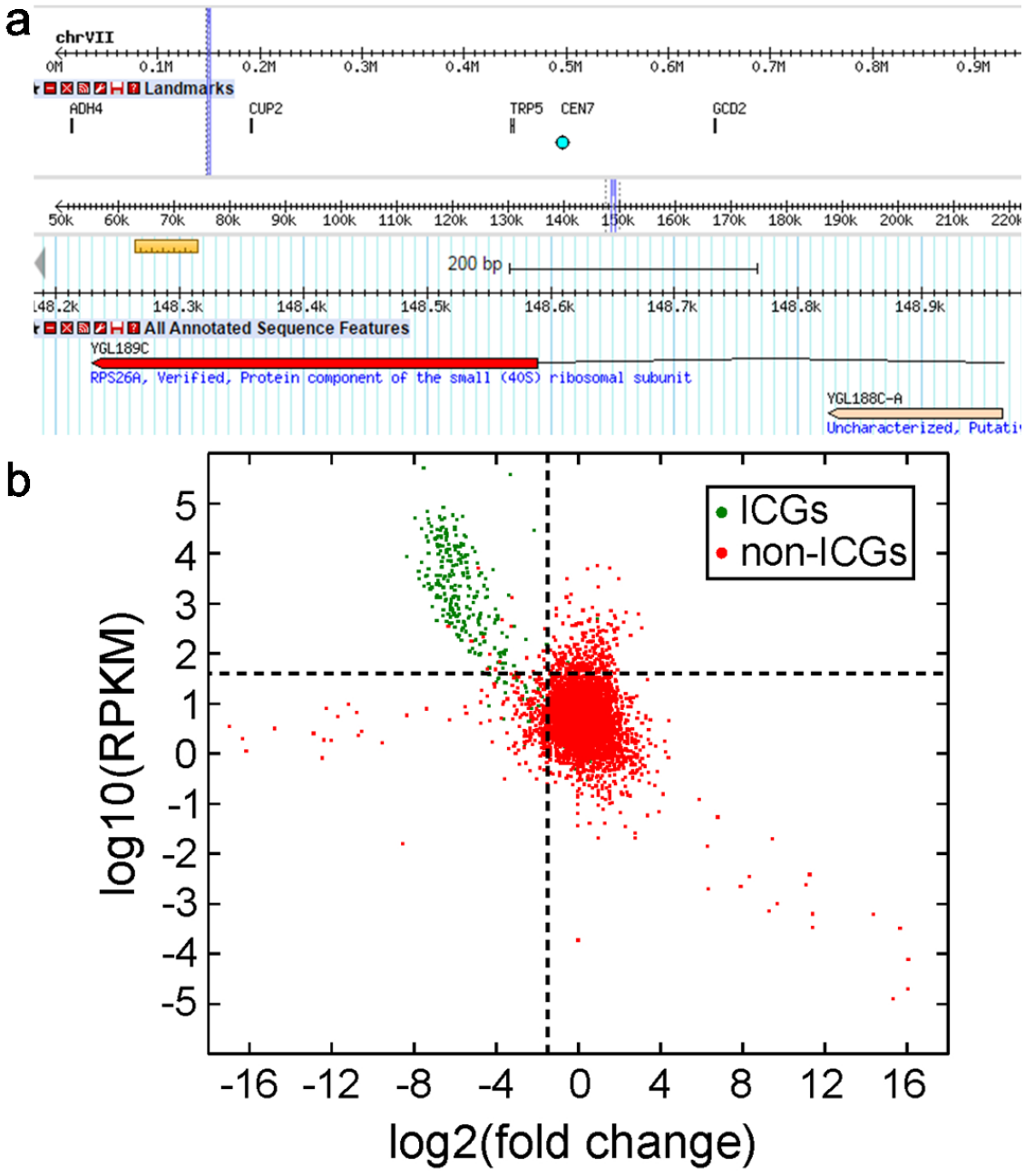



**Figure 5.9 Several highly abundant genes overlap with an intron-containing gene**
(a) A majority of the most highly abundant, non-intron genes overlap with known intron-containing genes. Shown is a screenshot of a non-intron gene (YGL188C-A, RPKM value of 3.89) overlapping with a highly abundant intron-containing gene (RPS26A, RPKM 4.43). (b) Scatter plot depicting the RPKM and fold change values relative to total extract RNA for ICGs and non-ICGs after removal of overlapping genes as in (a). Vertical and horizontal lines indicate the fold change and RKPM thresholds of -1.5 and 1.6, respectively.

| Gene ID | Low salt $B^{act}$ RPKM ($log_{10}$) | $log_2$(fold change) |
|---|---|---|
| YBR099C | 2.25 | -5.22 |
| YNL195C | 1.68 | -4.04 |
| YMR134W | 1.84 | -3.82 |
| YDL070W | 2.68 | -3.71 |
| YGR182C | 2.55 | -3.27 |
| YMR148W | 1.60 | -3.27 |
| YHR199C | 1.92 | -2.50 |
| YLR194C | 1.95 | -1.92 |
| YPR154W | 2.06 | -1.69 |
| YDL048C | 1.88 | -1.66 |
| YDR077W | 2.49 | -1.65 |
| YDR055W | 1.69 | -1.51 |
| YPL014W | 1.80 | -1.49 |

**Table 5.2 Top 13 non-ICGs**
Top non-ICGs showing significant association with the spliceosome (RPKM > 1.6 and fold change < -1.49)

gene of unknown function thought to share a promoter with the upstream *YNL194C* gene, a transcript with a predicted intron. Recent RNA-seq data detected *YNL195C* transcripts possessing the entire *YNL194C* gene as well as a shorter, spliced form possessing a small fraction of the 5' end of *YNL194C* (**Figure 5.11a**)[163]. Interestingly, RNA can be detected in the $B^{act}$ complex for both the *YNL195C* and *YNL194C* transcripts. YNL195C is not predicted to contain an intron and thus, in order to be present in the spliceosome and detected in our sample, must either be transcribed as a single unit with YNL194C ("bleed-through") or the predictions are incorrect and YNL195C does in fact contain an intron. The near equal abundance of both genes within the $B^{act}$ complex supports their transcription as a single unit and would explain the presence of YNL195C in the $B^{act}$ complex. Unfortunately, no reads were found in the intergenic region between the two genes as would be expected if these two genes are transcribed and spliced as a single transcript (**Figure 5.11b**)[163]. It is possible that this region of the RNA is subject to insufficient ligation or amplification during cDNA preparation resulting in the low read depth between the two genes. Alternatively, YNL195C may in fact contain an intron and become incorporated into the $B^{act}$ complex independently of YNL194C. Further experimentation will be required to fully understand YNL195C's detection at high levels within the spliceosome. Lastly, YJL206C was shown in the same publication to often act as a second exon with its upstream gene YJL205C (**Figure 5.12a**). YJL205C and YJL206C are often expressed as a single, long transcript that results in use of a 3'SS within the 5' end of YJL206C and removal of a majority of the JYL206C gene. Alternatively, both genes have separate promoters and thus can be expressed as separate transcripts. This would allow YJL205C to be recognized and spliced by the spliceosome at its canonical, internal splice sites. Interestingly, both full-length and individually expressed transcripts are supported by our data (**Figure 5.12b**). The number of

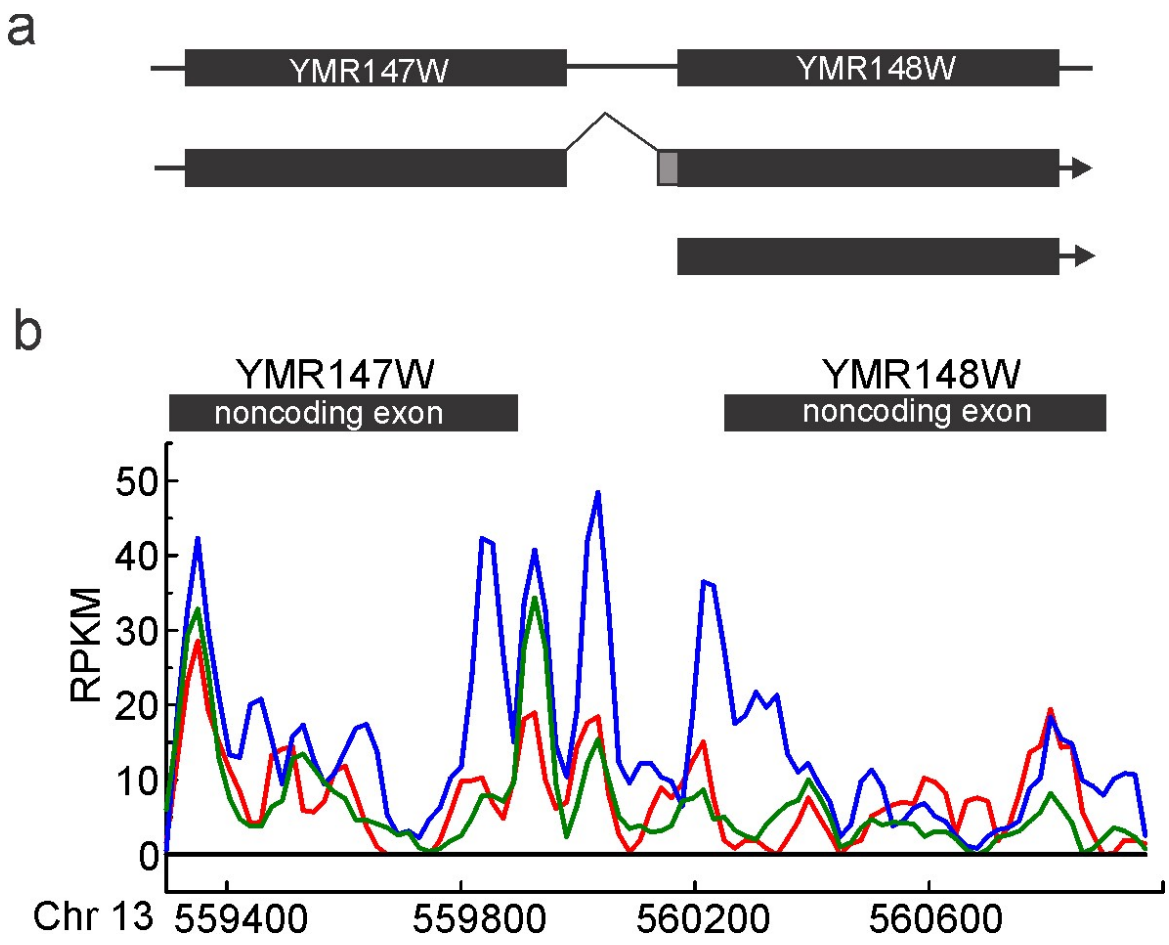



**Figure 5.10 YMR147W and YMR148W appear as a single transcript**
(a) Previously predicted spliced products from Miura et al. 2006.
(b) Mapped reads from YMR147W and YMR148W in the $B^{act}$ complex washed with either 150 mM (green and red curves) or 500 mM NaCl (blue curve) during purification.

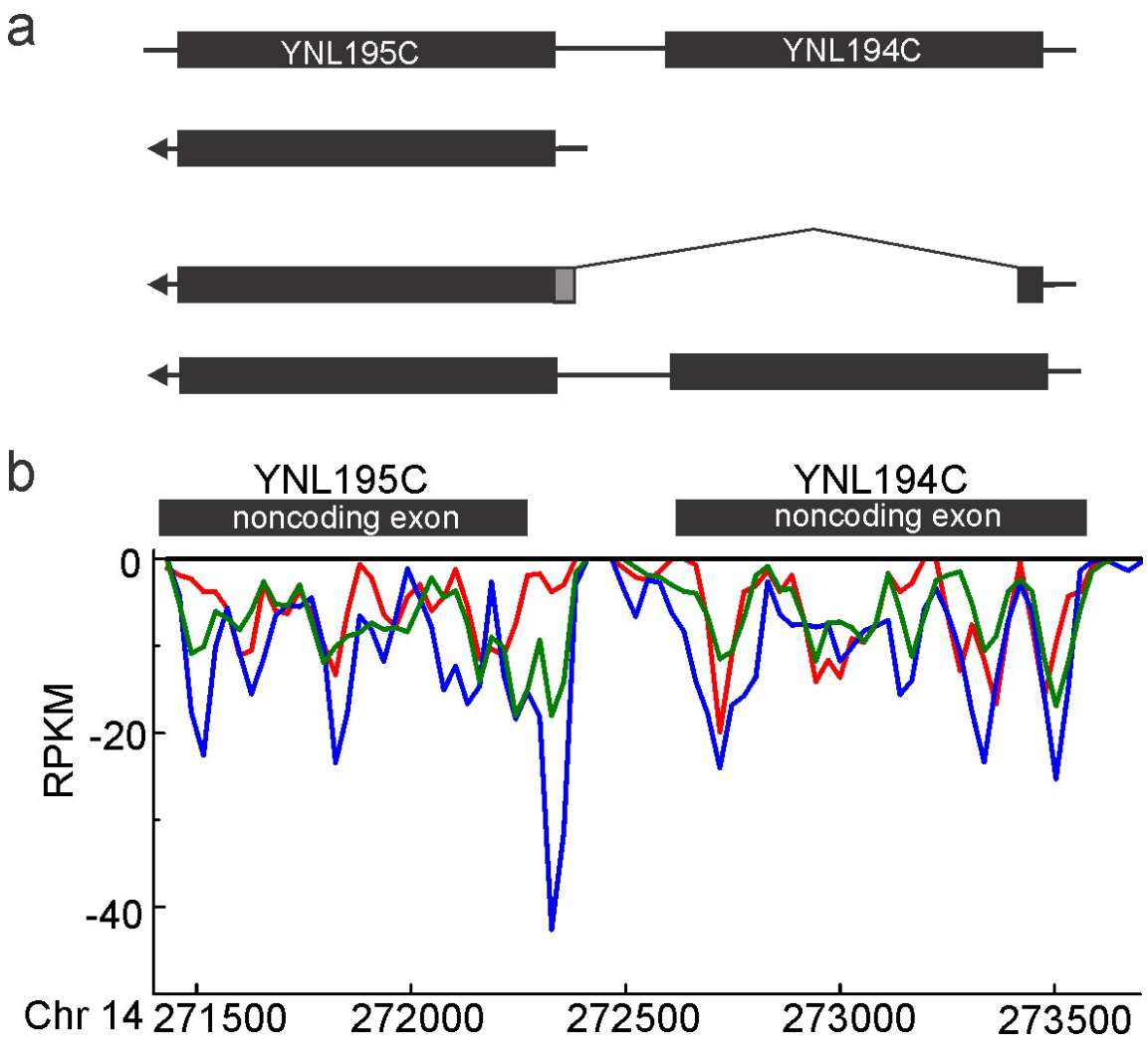



**Figure 5.11 YNL194C and YNL195C comprise a single transcription unit**
(a) Previously predicted spliced products from Miura et al. 2006.
(b) Mapped reads from YNL195C and YNL194C in the $B^{act}$ complex washed with either 150 mM (green and red curves) or 500 mM NaCl (blue curve) during purification.

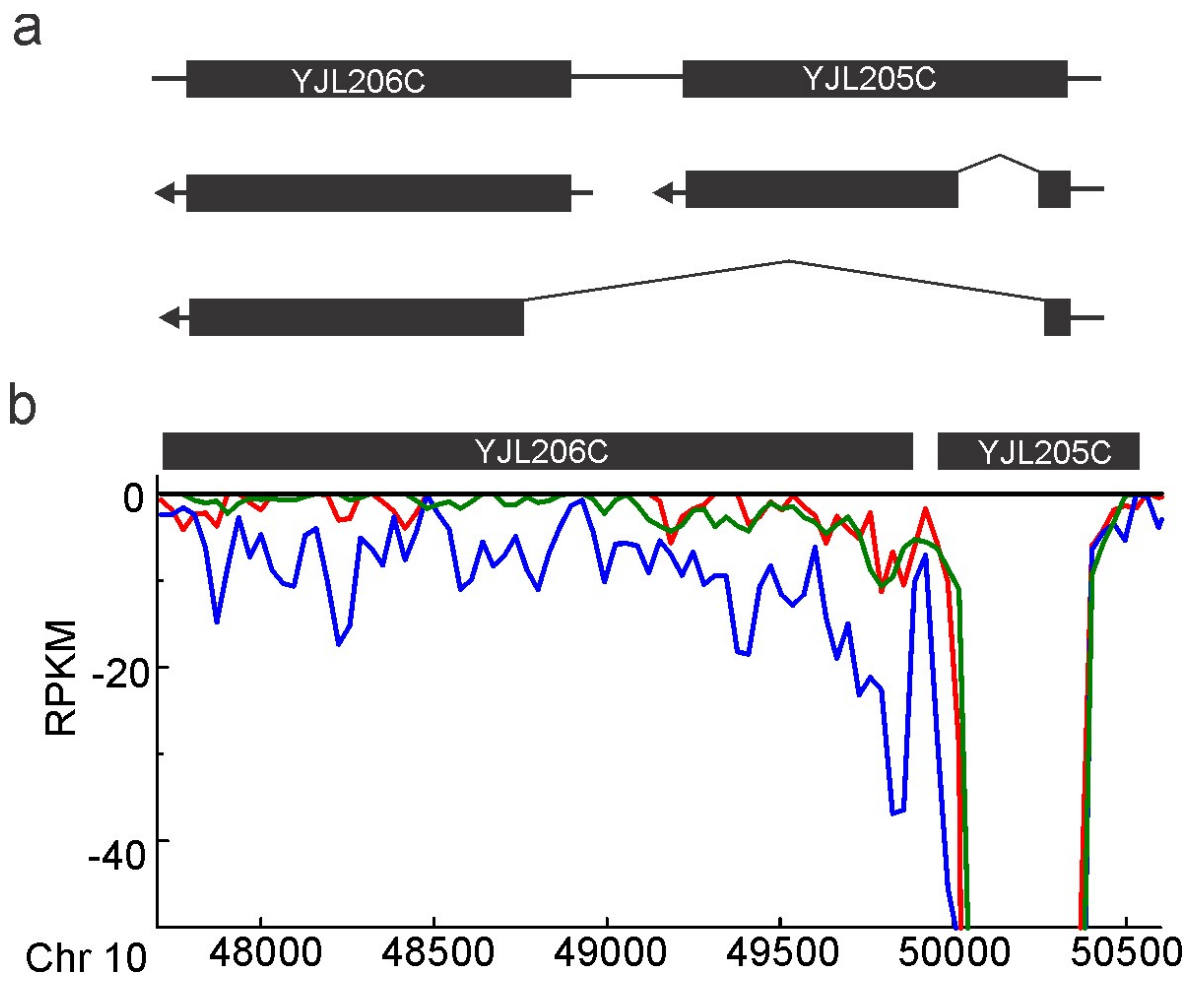



**Figure 5.12 YJL206C can be spliced to its upstream gene**
(a) Previously predicted spliced products from Miura et al. 2006.
(b) Mapped reads from JYL206C and YJL205C in the $B^{act}$ complex washed with either 150 mM (green and red curves) or 500 mM NaCl (blue curve) during purification.

mapped reads within the YJL205C gene is nearly 100-times greater than the YJL206C gene indicating the separate YJL205C RNA is the primary transcript within the Bact complex. There is, however, a significant amount of reads mapped to YJL206C that also show connectivity to the upstream YJL205C gene. Therefore, although to a much lesser extent, YJL206C might be expressed as a single transcript with YJL205C and thus be incorporated into the $B^{act}$ complex for splicing.

YMR134W is a verified protein of unknown function involved in ergosterol biosynthesis and is located just downstream of a known intron-containing gene YMR133W. YMR133W is a meiotic gene and thus is present in very low levels in the $B^{act}$ complex. The very 3' end, however, shows a large number of mapped reads that start at the YMR133W intron and continue through the entire downstream YMR134W gene (**Figure 5.13**). Often, meiotic genes such as YMR133W have promoters that drive antisense transcription near their 3' ends[170,171]. It thus could be that a 3' promoter present just upstream of YMR133W's intron becomes activated under our growth conditions and drives transcription in the other direction so that the very 3'end positioned YMR133W intron is transcribed and some of the transcripts extend into YMR134W. As such, the YMR134W transcript would appear to have a 5'UTR intron allowing for recognition and incorporation into the spliceosomal $B^{act}$ complex.

It was recently discovered that the spliceosome can participate in regulation of RNA and protein expression through a pathway known as Spliceosome-Mediated Decay (SMD)[164]. SMD involves the downregulation of genes not known to contain introns but that do contain canonical splice sites. Genes subject to SMD remain primarily unspliced and full-length in the yeast cell in order to yield the functional, full-length protein. However, these genes can be recognized by the spliceosome and proceed through one or both steps of splicing. By doing so, a truncated

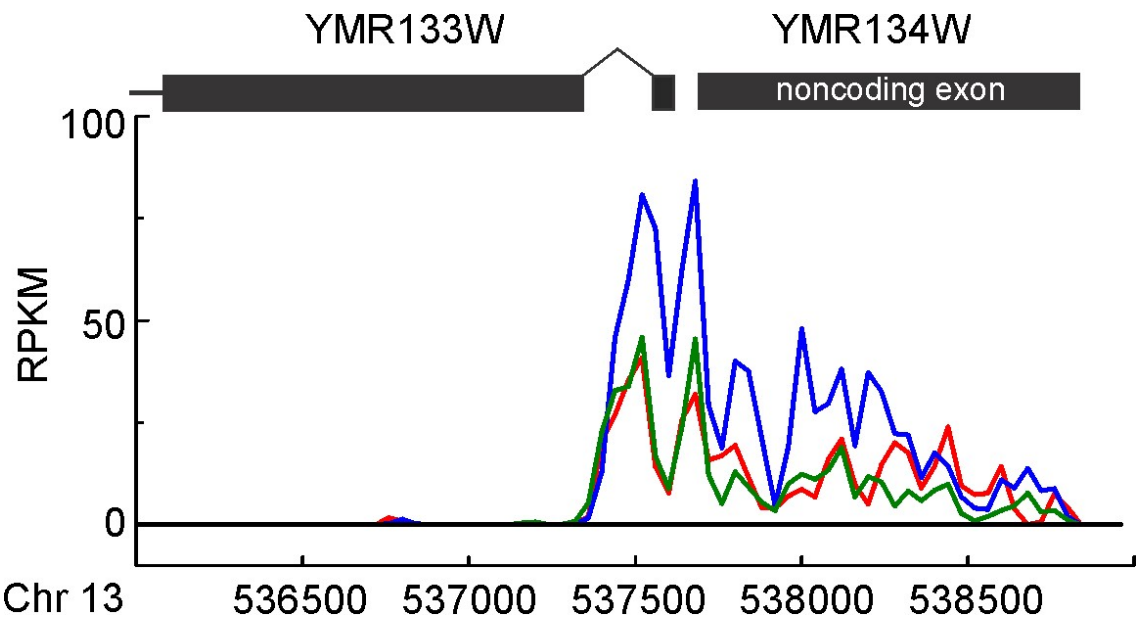



**Figure 5.13 YMR134W appears to contain a 5'UTR intron**
Shown are the mapped reads for the regions encoding the YMR133W and YMR134W genes. The $B^{act}$ complex was washed with either 150 mM (green and red curves) or 500 mM NaCl (blue curve) during purification.

transcript is produced that is unusually unstable and thus is rapidly degraded by the nuclear RNA surveillance machinery. Such a mechanism allows for the maintenance of proper transcript levels for genes such as BDF2, a protein known to be toxic at elevated levels. Interestingly, BDF2 (YDL070W) is one of the 13 new intron-containing genes detected in our Bact complex indicating that our assay is capable of detecting spliceosome-mediated decay targets (**Figure 5.14**). Additionally, there is a significant peak in the number of mapped reads near the location of the proposed 5'SS. Peaks in the number of reads often occur near intronic regions due to protection from degradation by the spliceosome. Our data grant further support towards the existence of the SMD pathway and to BDF2 being subject to this type of regulation.

YBR099C is a dubious or putative open reading frame unlikely to encode a functional protein based on comparative sequence data. This gene was not detected as being spliced or even transcribed in previous work that sought to determine all transcription start sites and splice sites in yeast[163,164]. Like YMR148W, YNL195C, and YJL206C, YBR099C is directly downstream of a known intron-containing gene FES1 and appears to show mapped reads that connect FES1 with the YBR099C transcript (**Figure 5.15**). If FES1 possesses a weak transcription terminator, the detection of YBR099C could simply be a result of transcript bleed-through. However, the large number of reads near the middle of YBR099C suggests that the RNA is strongly protected from degradation by the spliceosome and thus that YBR099C might contain an intron-like sequence in this region of the gene. One explanation could be that the FES1 gene is subject to an abortive splicing event such as 3' end cleavage. Abortive splicing is used to make the mature 3' end of *S. pombe* and some fungal telomerase RNAs[172]. Additionally, SMD can be an abortive splicing event used to reduce the levels of particular RNA transcripts. Another explanation is that, like YMR148W, YNL195C, and YJL206C, YBR099C may contain an alternative 3'SS for splicing

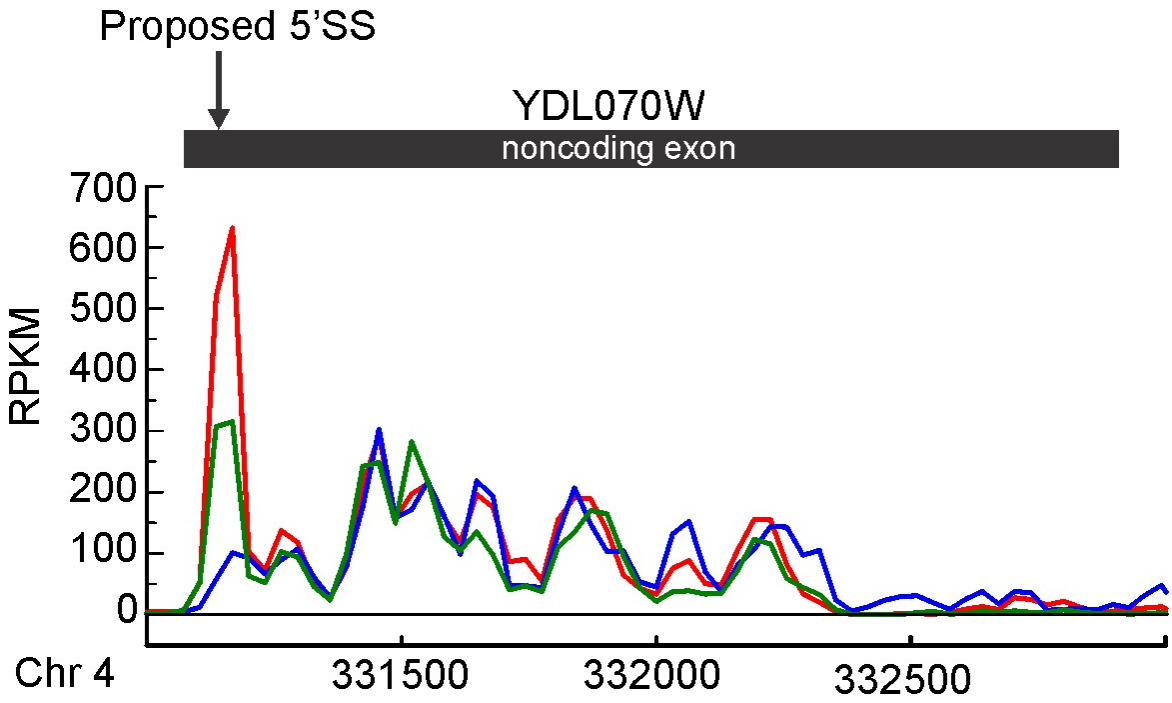



**Figure 5.14 A SMD target YDL070W is detected in the $B^{act}$ complex**
YDL070W (BDF2), a SMD target, is detected in the $B^{act}$ complex with a significant peak near the proposed 5'SS. The $B^{act}$ complex was washed with either 150 mM (green and red curves) or 500 mM NaCl (blue curve) during purification.

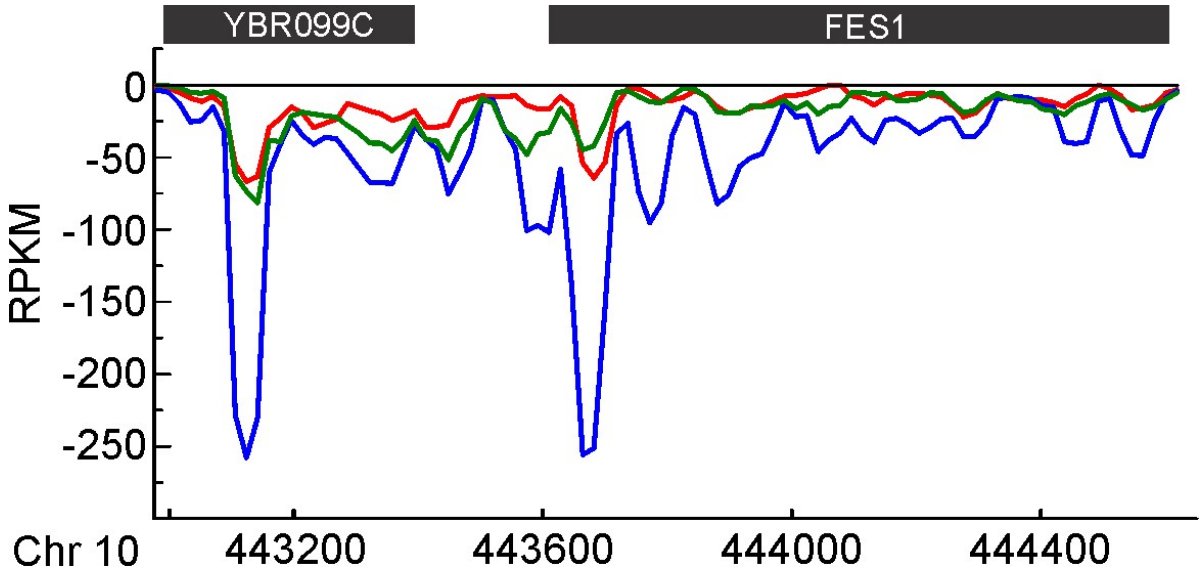



**Figure 5.15 FES1 may undergo an abortive splicing event**
FES1 was previously-predicted to contain an intron near the peak in mapped reads near its 5' end. YBR099C, however, is also detected in high abundance in the $B^{act}$ complex and shows a large peak in mapped reads in the middle of the gene indicating protection by the spliceosome. The $B^{act}$ complex was washed with either 150 mM (green and red curves) or 500 mM NaCl (blue curve) during purification.

with FES1. Attachment of additional protein sequence to FES1 may change the function or localization pattern of FES1 as is thought to be the case for YNL239W[173]. Either scenario would explain the presence of a strong peak in the mapped reads downstream of the FES1 gene and requires further experimentation to validate the finding.

Finally, genes YDL048C, YDR077W, and YDR055W may also be subject to regulation by SMD. All three RNAs show significant mapped reads across the entire gene and even appear to have peaks near their 5' ends, indicative of protection from degradation by the spliceosome (**Figure 5.16a-c**). It was found that a key feature of the SMD pathway is the recruitment of the spliceosome to the BDF2 transcript by its paralog BDF1[164]. BDF1 is known to have a connection of spliceosome recruitment to pre-mRNA transcripts as deletion of BDF1 reduces splicing of a large subset of intron-containing transcripts[174]. However, aside from BDF2, recruitment of the spliceosome to genes lacking introns has not been further investigated. In addition, it is estimated that up to 1% of yeast intronless genes may be subject to regulation by the SMD pathway[164]. Therefore, it is entirely conceivable that YDL048C, YDR077W, and YDR055W transcripts may be recruited to the spliceosome by BDF1 or some other unknown factor for regulation of cellular protein levels by the SMD pathway.

**5.4 Discussion**

Here we have isolated the *in vivo*-assembled yeast activated spliceosome ($B^{act}$ complex) containing nearly all known pre-mRNA substrates for future RNA secondary structure analysis using SHAPE-MaP. The current understanding is that the primary components of the spliceosome are the five snRNA molecules and all the associated protein factors. The snRNP complexes assembly upon a pre-mRNA substrate in order to carry out both steps of splicing and produce the mature RNA coding sequence that is translated into protein. Unfortunately, the pre-

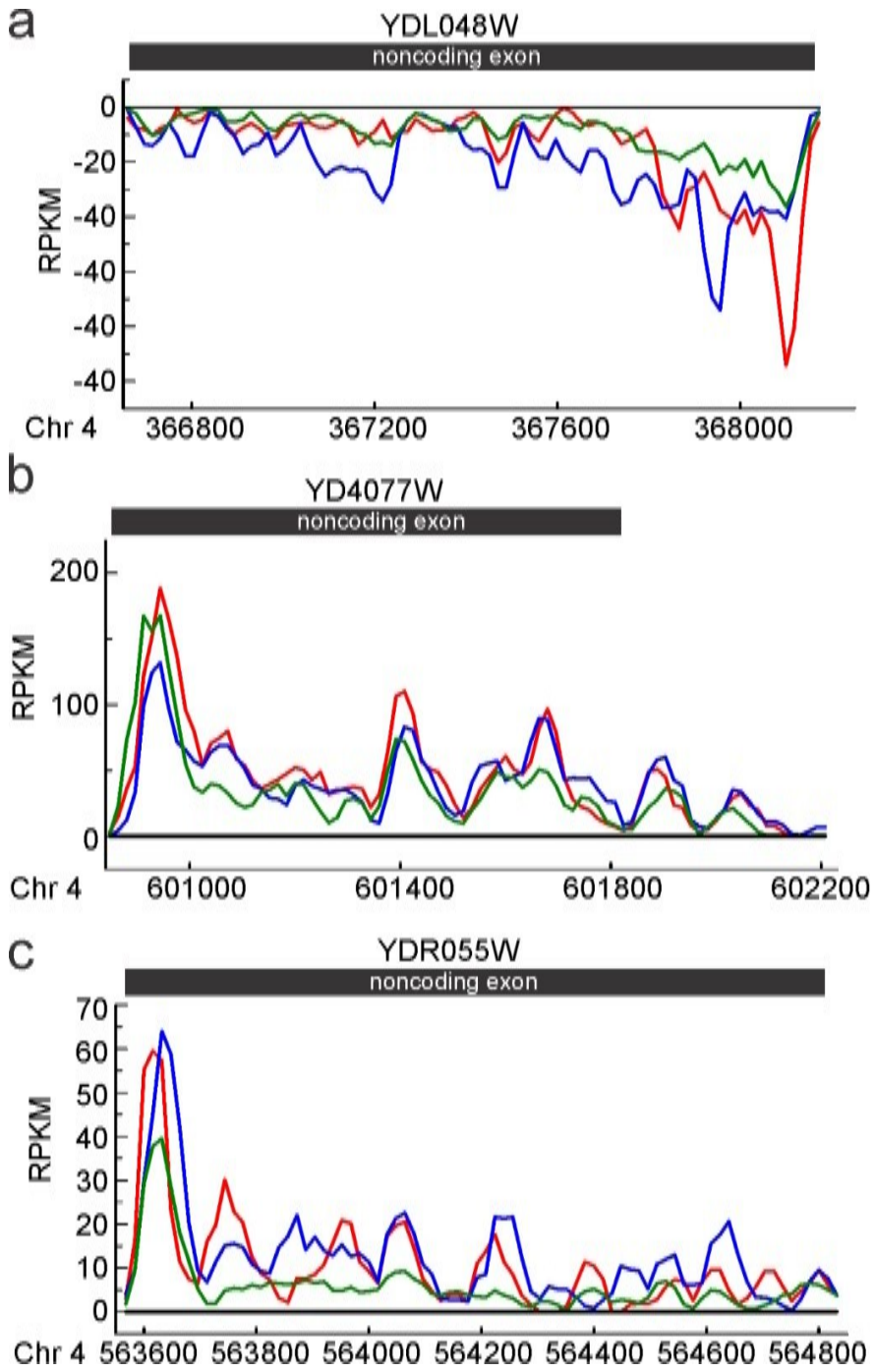



**Figure 5.16 Several candidate pre-mRNA targets could be under regulation by the SMD pathway**

Candidate pre-mRNA substrates possessing strong signals at the 5' end and significant reads over the entire length of the gene. The $B^{act}$ complex was washed with either 150 mM (green and red curves) or 500 mM NaCl (blue curve) during purification.

mRNA is all too often regarded as just a substrate for the spliceosome and as a result ignored as an active participant in splicing catalysis and regulation. There is, however, increasing evidence that the pre-mRNA substrate and specifically its secondary structure can influence splicing and splicing efficiency[154-157] supporting the hypothesis that intron secondary structure does play a functional role in splicing. Therefore, our goal is to delineate the functional impact of intron secondary structure on splicing activity by determining the secondary structure of all pre-mRNA substrates within *in vivo*-assembled spliceosomes.

Proper isolation of the yeast $B^{act}$ complex was achieved using a temperature-sensitive strain carrying a TAP-tagged Cef1 protein (Prp2-1,Cef1-TAP) and confirmed by RT-PCR and Northern blot analysis. To first confirm the presence of known pre-mRNA targets and possibly identify new splicing substrates, the RNA from the complex was isolated and submitted for RNA sequencing (RNA-seq). To ensure enrichment of RNA relative to levels in a total RNA control, RNA from the yeast whole cell extract from which the complex was isolated was also sequenced. Additionally, spliceosomes were either washed with a low salt (150 mM NaCl, LS) or high salt (500 mM NaCl, HS) buffer during purification to identify any RNA substrates loosely associated with the spliceosome and that are thus most likely not splicing targets. Computational analysis of the data revealed significant enrichment of over 90% of the known pre-mRNA targets that would be predicted to be present in the $B^{act}$ complex isolated during vegetative growth. Nearly half of the known snoRNAs were also detected in the LS-purified complex but were nearly completely removed in the HS-purified complex indicating loose association probably through base pair interactions with the co-purified ribosomal RNA. Upon removal of overlapping gene reads, the $B^{act}$ complex was found to contain an addition 13 genes that were previously not predicted to contain an intron. Remarkably, three of these genes were previously predicted to be expressed as

a single transcript with the upstream intron-containing gene (YMR148W, YNL195C, YJL206C) and involved in an alternative splice site usage mechanism[163]. Additionally, the isolated $B^{act}$ complex contains significant levels of an RNA recently shown to be subject to regulation by the spliceosome-mediated decay (SMD) pathway (YDL070W), signifying that our assay is capable of detecting and potentially predicting SMD targets. As such, three of the remaining RNA transcripts (YDL048C, YDR077W, YDR055W) show significant enrichment in the $B^{act}$ complex and also possess a large number of reads near their 5' ends, similar to YDL070W. It is possible that these genes possess canonical splice signals within the 5' ends of the RNA and are also subject to regulation by the SMD pathway.

With a high-confidence list of known pre-mRNA substrates found within the $B^{act}$ complex, we next plan to carry out RNA structure prediction using SHAPE-MaP analysis[63] in collaboration with the Laederach lab at the University of North Carolina. HS-purified Bact complexes will be prepared and incubated in the presence or absence of 1M7 SHAPE reagent[175]. Additionally, a denatured control will be included in order to allow for efficient normalization of the SHAPE reactivity data. Specific secondary structure features will then be identified for each pre-mRNA substrate. We will keep track of the location of these structural features relative to the splice sites, the linear distance between splice sites as a result of the structures, as well as the splicing efficiency of each substrate in order to identify pre-mRNA structural features that either promote or delay splicing. If a correlation is found between specific structural features and splicing efficiency, we will take this as support of our hypothesis that the pre-mRNA substrate can actively participate in splicing through intrinsic adoption of an optimal secondary structure.

Following thorough SHAPE-MaP analysis, we will also continue to identify potential explanations for incorporation into the spliceosome of RNA transcripts for which there is no

predicted intron. Similar to previous studies, we will test for regulation via the SMD pathway using Northern blot and RT-PCR analysis to identify full length and potentially spliced RNA transcripts. Potential spliced products will be enriched through use of mutant strains of yeast in which either the debranching enzyme or exosome is inactivated. Finally, we will test if BDF1 has an impact on the splicing levels of these candidate RNA targets through use of a BDF1 deletion strain.

# CHAPTER 6: Conclusions and Outlook

## 6.1 Conclusions

Identification and characterization of protein and RNA structures is fundamental to achieving a thorough and complete understanding of all cellular processes. Oftentimes, specific structures or folds are more conserved than the sequence of building blocks that make up a macromolecule. Therefore, structural, as well as dynamic, information has the great potential to provide a more meaningful way to characterize cellular function than sequence conservation and similarity alone. This is especially true for RNA, with its central roles in transcription initiation, elongation, and termination, pre-mRNA splicing, translation, and retroviral infection of eukaryotic cells. RNA molecules typically possess a diverse array of complex secondary and tertiary structures that give rise to intricate and fluid 3-dimensional architectures that allow for specific interactions with other nucleic acids, proteins, and small molecules.

Current work in structural biology primarily utilizes X-ray crystallography and electron microscopy (EM) for the characterization of large and small biomolecular machines. These techniques typically use specific mutations or chemical modifications in or drugs against the enzyme of interest or its substrate to stall the biomolecule in a pre- or post-catalytic conformation. Depending upon the resolution, detailed structural analysis can reveal vital information about substrate binding or product release, allowing for an educated guess about the predicted mechanistic pathway that powers the biological function at the catalytic sites. As a result, a significant attempt has been made in the last 15 years to identify the structure of the spliceosome and its components using these techniques[36,176]. For example, low resolution

structures have been determined for the human B and C complexes[177], several individual snRNPs[30,178-180], and a number of crucial proteins known to be in the heart of the spliceosome during the catalytic steps of splicing[181]. Only recently have high resolution pictures of key components of the spliceosome, such as Prp8, become available[182]. Prp8 is a key component of the U5 snRNP showing extensive crosslinks to the RNA catalytic core of the spliceosome. Through crystallization of Prp8, the authors revealed a structure of the 'heart of the spliceosome' showing great similarity to a bacterial group II intron reverse transcriptase. Interestingly, known suppressors of splice-site mutations mapped to a region of Prp8 large enough to accommodate the catalytic core of group II intron RNA. These higher resolution images of the individual components have aided in producing better estimations of the intact spliceosome by mapping these structures into the low resolution structures of intact spliceosomes. Furthermore, work has begun to map the exact location of particular regions of the pre-mRNA substrate within the spliceosome's active site[183,184]. Unfortunately, our knowledge about the structure-function relationship is limited with these before and after images of biological machines such as the spliceosome. EM and crystal structures provide great starting points but reveal very little about the dynamic mechanisms associated with most biological machines. Only upon meticulously working to understand how structural dynamics produce a given functional outcome can we properly fine-tune and manipulate macromolecular sequence, and therefore structure, to obtain a desired effect. In this thesis, we have started to address these challenges utilizing smFRET and deep sequencing-mediated structure prediction (SHAPE-MaP) to directly investigate the time series of pre-mRNA structures and the dynamics between them required to efficiently catalyze splicing.

**Single-Molecule Pull-down FRET (SiMPull-FRET) to dissect the mechanism of first-step splicing catalysis**

Single-molecule fluorescence resonance energy transfer (smFRET) has recently emerged as an invaluable technique capable of monitoring pre-mRNA splice-site proximity throughout splicing assembly and catalysis[53]. This approach revealed a large set of reversible time- and ATP-dependent conformational dynamics as the spliceosome assembles upon a fluorescent substrate and carries out both steps of splicing. While real-time structural information like this is crucial to our understanding of the splicing mechanism, the associated multiple FRET states with varying kinetics are increasingly difficult to understand and assign to particular splicing complexes.

As a means to enrich for and investigate a particular step of the splicing cycle, we developed a technique that couples purification of specific splicing complexes with smFRET in a method we termed SiMPull-FRET[55]. Utilizing a Prp2-1,Cef1-TAP yeast strain and a fluorescent pre-mRNA containing donor and acceptor fluorophores near the BP and 5'SS, respectively, SiMPull-FRET allowed for the isolation and investigation of the protein-dependent pre-mRNA transitions associated with the first chemical step of splicing. smFRET analysis of the purified $B^{act}$ complex revealed a static, unchanging low FRET state indicating a large separation in distance of the reactive 5'SS and BP. Such a conformation is thought to be induced by the SF3 complex, a small protein complex known to bind the BS sequence in the $B^{act}$ complex and prevent premature attack of the BS adenosine on the 5'SS[39]. Addition of Prp2, Spp2, and ATP to the purified $B^{act}$ complex, conditions shown to destabilize association of the SF3 complex, rearranged the substrate to reversibly explore conformations with a proximal 5'SS and BP that are capable of carrying out low levels of first step splicing. Only upon addition of Cwc25, however, does this equilibrium become strongly biased towards the proximal conformation,

promoting efficient first-step splicing. Such a mechanism is reminiscent of the biased Brownian ratcheting mechanism utilized by the ribosome in which a helicase unlocks thermal fluctuations that are subsequently rectified by a cofactor 'pawl'. We have, therefore, not only discovered a mechanism used by the spliceosome to achieve efficient first-step splicing, we have also shown that smFRET, coupled with single-molecule pull-down, is an invaluable tool for the investigation of the spliceosome and other RNA-based machines.

**Assigning conformational dynamics to specific splicing complexes using Single-Molecule Cluster Analysis (SiMCAn)**

Many of the dynamic processes required for the proper assembly, catalytic activation, and disassembly of the spliceosome as it acts on its pre-mRNA substrate remain poorly understood. The enrichment and purification of all spliceosomal complexes for study by SiMPull-FRET, as utilized to study the Prp2 and Cwc25-mediated enhancement of the first step of splicing[55], would be tedious and time-consuming. We have therefore implemented the use of several, well-established biochemical stalls to enrich for specific splicing complexes throughout spliceosome assembly and catalysis. Through incubation of an immobilized WT or 3'SS mutant yeast substrate containing the same donor and acceptor fluorescent dyes near the BP and 5'SS with yeast whole-cell extract (WCE) containing the mutation of interest, we were able to monitor the pre-mRNA dynamics associated with spliceosome assembly up to a defined endpoint in the splicing cycle. In order to dissect the manifold conformational dynamics of the pre-mRNA in each of the splicing block conditions we developed Single Molecules Cluster Analysis (SiMCAn), a bioinformatics clustering tool capable of grouping and sorting single-molecule FRET data based on common dynamic behavior. Through the implementation of a second round of clustering that grouped clusters based on their occupancy across the set of experimental

conditions, SiMCAn was able to identify signature conformations and dynamic behaviors of multiple ATP-dependent intermediates. In addition, it identified a conformation adopted late in splicing by a 3'SS mutant substrate in which the 5'SS and BS become stably removed from one another, invoking a mechanism for substrate proofreading.

We have, therefore, developed an alternative method using simple biochemical stalls and bioinformatics clustering analysis capable of assigning FRET states and dynamics to specific splicing complexes that does not require purification of individual splicing complexes. The SiMCAn method presents a novel framework for interpreting complex single molecule behaviors that should prove widely useful for the comprehensive analysis of a plethora of dynamic cellular machines.

**Assigning RNA structural pathways to real-time, FRET-based conformational dynamics**

Single-molecule FRET has allowed for the real-time observation of changes in RNA, protein, and complex structure that are beyond the capabilities of conventional structure prediction analysis. Such experimentation revealed an unprecedented glimpse into the heart of both simple and complex RNA machines, such as the spliceosome, revealing specific details about the changes in structure required to carry out a function. While these strategies have exposed much about common and often complex mechanisms, there is a gap in our ability to reliably translate time-resolved smFRET data into a temporal sequence ('movie') of secondary structures that an RNA molecule adopts throughout smFRET observation.

In an effort to circumvent these limitations, we have developed FRETtranslator, an RNA structure prediction software capable of confidently predicting RNA secondary structures through incorporation of both smFRET and biochemical footprinting data. To first optimize and

validate our approach, we applied FRETtranslator to smFRET data gathered from a short, 76-nucleotide long RNA containing the Cy3-Cy5 FRET pair near the base of a highly variable region of the RNA. Intriguingly, FRETtranslator predicted several secondary structures that efficiently mirror the FRET states and kinetics observed during smFRET analysis. Specifically, FRETtranslator identified a compact structure describing the dominant high FRET conformation as the most common observed structure. This structure shows great similarity to an alternative high FRET structure most common to molecules exhibiting a slightly higher, 0.85 FRET state. While one specific path was observed for substrate unfolding to a low FRET state, two populations of behaviors were observed for substrate refolding as determined through kinetic analysis as well as by FRETtranslator. These results indicate that FRETtranslator can efficiently predict RNA secondary structures from a series of interconverting FRET states containing varying transition kinetics.

To further apply FRETtranslator to a biologically relevant system, we collected smFRET and footprinting data for the full length Ubc4 splicing substrate. We next plan to use FRETtranslator to predict the most likely secondary structures of Ubc4 in buffer as well as in several splicing complexes with the hope of creating a sequence of secondary structures adopted by single pre-mRNA molecules during spliceosome assembly and catalysis.

**Identifying a structure-function relationship in yeast pre-mRNA substrates**

In several initial studies, it was found that the efficiently spliced Ubc4 intron exhibits significant secondary structure that places the points of first and second step chemistry much closer together than would be expected from their linear sequence distance. This would suggest that the secondary structure is such that it places the reactive sites in close proximity before the association of any protein or RNA components. Such an orientation shows great similarity to that

of group II introns that are capable of splicing in the absence of all protein cofactors and supports a model where the intron plays an active role in splicing.

To further test the hypothesis that specific intron secondary structures dominate in spliceosomal cycle intermediates, we have isolated the *in vivo* assembled, yeast $B^{act}$ complex containing over 250 of the known yeast pre-mRNA substrates for RNA structure analysis using selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP). Northern blot and RT-PCR analysis confirmed the presence of only the expected snRNAs (U2, U5, and U6) as well as predominantly unspliced Act1 pre-mRNA. We further confirmed the purity of the complex through RNA-seq analysis of the RNA contained in the complex, revealing a significant enrichment of genes previously predicted to contain introns. Additionally, we purified the $B^{act}$ complex under high salt conditions to further remove non-specifically associated RNA such as snoRNAs. Interestingly, we were able to show that our assay is capable of detecting the spliceosome-mediated decay (SMD) target YDL070W, as well as several bicistronic genes recently found to undergo alternative splicing events. Furthermore, we identified a number of new potential intron-containing genes or SMD targets within the complex.

We next plan to perform SHAPE-MaP analysis on the confirmed pre-mRNA targets within the spliceosome in order to determine the secondary structure of all RNA within the complex. We will then cluster the consensus secondary structures found among all introns to test the hypothesis that certain intrinsic secondary structures correlate with high splicing efficiency.

### 6.2 Outlook

Proper RNA folding is crucial to cellular function. In doing so distinct regions of an RNA molecule with a potentially large separation in their primary sequence can now interact to provide structural integrity, increase the accessibility of a protein or RNA binding site, or allow

proper formation of a catalytic active site. In addition to potentially changing a catalytic residue, alterations in nucleic acids (mutations) can also lead to changes in folding and thus prevent proper cellular function. Such could be the case for the up to 50% of all mutations that lead to human disease that act through disruption of the splicing code[153,185-188]. These mutations often result in disruption of the exon-intron boundaries and prevent recognition and splicing by the splicing machinery leading to formation of aberrant mRNAs that are unstable or lead to formation of defective protein isoforms. Interestingly, over 400 intronic single-nucleotide variations (SNVs) that are more than 30 nucleotides from any splice site were recently discovered that induce changes in splicing patterns far more severe than common variants[189]. This same study revealed thousands of exonic mutations and tens of thousands of other disease-causing mutations that have a great potential to alter splicing. In order to fully understand how these mutations affect splicing and the downstream function requires a thorough understanding of how the mutations affect RNA structure through direct, single-molecule and single-nucleotide resolution visualization of these structures.

In recent years, single molecule FRET has begun to shed light on mechanisms used by the spliceosome to achieve high catalytic efficiency and specificity. In particular, the RNA helicases found throughout the splicing cycle are thought to utilize the energy from ATP hydrolysis to bind, unwind, and release RNA, thus remodeling inter- and intra-molecular RNA structures and dissociating associated proteins. The study of RNA helicases has primarily been performed in highly purified *in vitro* systems leaving many unanswered questions regarding the precise kinetics and mechanism of DExD/H-box RNA helicase function within the complex RNP machines they are most typically known to operate. In this thesis, we describe our pioneering work on the processive DEAH-box helicase Prp2 and its involvement in the first step of splicing

using SiMPull-FRET. A rather straightforward extension of this is the investigation of Prp16's ATP-dependent and –independent roles before, during, and after the first step of splicing using SiMPull-FRET. Prp16 is a second-step factor known to catalyze the ATP-dependent removal of Cwc25 after the first step of splicing, allowing for the formation of the second step conformation[85]. Prp16 is also thought to facilitate a kinetic proofreading mechanism wherein Prp16 acts as a timer to ensure that suboptimal branchsites are deprived of the possibility to react[127]. If splicing is slow, ATP hydrolysis by Prp16 will result in the premature removal of Cwc25 prior to catalysis and thus discard of the substrate. Interestingly, two separate studies have found that Prp16 is capable of assisting with the first step of splicing in both an ATP dependent and independent manner[41,190]. In the first study, it was found that a mutated branchsite (A→C) normally unable to proceed through the first step of splicing could achieve first step catalysis upon incubation with an ATPase-deficient mutant of Prp16[41]. It was found that the mutant Prp16 facilitates the stabilization of Cwc25 with the branchsite, resulting in stabilization of a proximal BS-5'SS capable of achieving efficient first-step splicing. Incubation with WT Prp16, however, results in discard of the substrate. In contrast, the second publication discovered that if a deoxyribose branchsite adenosine is encountered, first-step splicing can be permitted through addition of WT Prp16 and ATP but not through use of a mutant Prp16[190]. Utilizing primer extension, the authors discovered that Prp16 can actually unwind the U2-BS interaction allowing for alternative BS adenosine selection and first-step catalysis. Unfortunately, neither of these studies was able to report on the mechanistic relationship between Prp16 action and pre-mRNA conformational changes, information that can easily be achieved via SiMPull-FRET, and thus could not determine the proofreading mechanism of Prp16.

In order to advance the use of smFRET and SiMPull-FRET for the study of splicing, future studies will be required to extend these approaches to the labeling of snRNAs and numerous proteins within the spliceosome. The five snRNAs work together to efficiently recognize and hand off the pre-mRNA splice sites during spliceosome assembly. Several studies have revealed discrete changes in snRNA structure at multiple stages of assembly and catalysis that are thought to be required for splicing. Having the ability to site-specifically incorporate fluorescent dyes into the snRNAs for visualization using smFRET would greatly improve our understanding of the roles snRNAs have in splicing. We have already demonstrated successful depletion and reconstitution of yeast splicing extracts with labeled U2 snRNA and observed significant preliminary binding in the presence of extract and ATP, conditions that allow complete assembly of all splicing components on the labeled pre-mRNA (**Figure 6.1b,c**). Such a setup allows for observation of snRNA assembly, but the fluorophores are not yet in FRET distance to provide additional information about snRNA-pre-mRNA dynamics. This is primarily due to a lack of available methods with which to internally label snRNA. The internal modification and labeling of RNA is greatly limited to short (<140 nt) RNAs due to our inability to efficiently synthesis RNA. Even so, synthesis and internal labeling of these short RNAs still requires synthesis of two RNA pieces that can be labeled and ligated as described in this thesis. One alternative for internally labeling RNA is the use of terbium-assisted deoxyribozymes[191]. This approach has already been used for the efficient labeling of *in vitro* transcribed U6 snRNA but should be amenable to either the U2 or U5 snRNAs. Once labeled, incorporation of the modified snRNA into yeast splicing complexes can be achieved through well-established snRNA depletion and reconstitution methods (**Figure 6.1a**)[192]. One immediate application would be to label the U2 BS-interacting region and incorporate the snRNA, along with BS-labeled dA Ubc4,

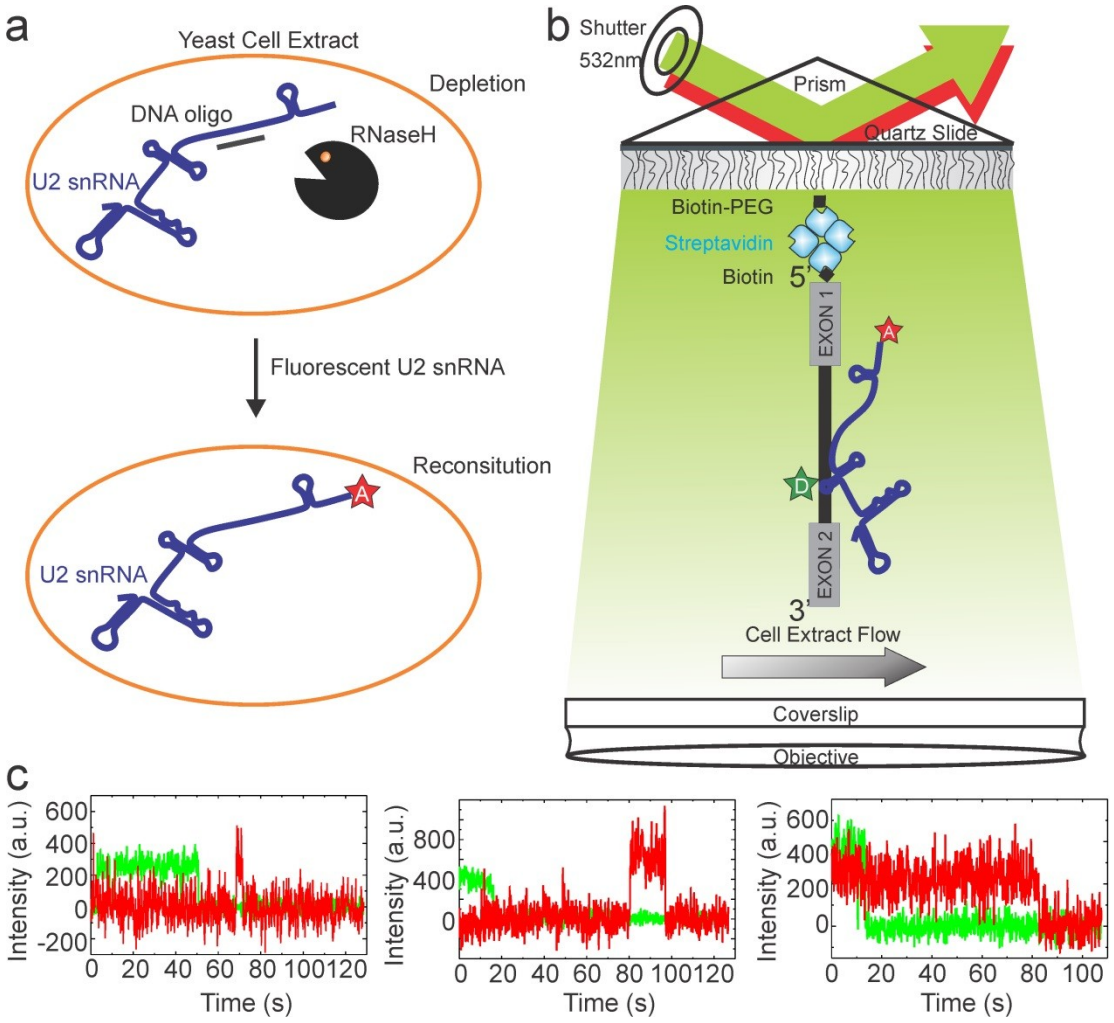



**Figure 6.1 Labeled U2 snRNA assembles on immobilized pre-mRNA**
(a) General protocol for the depletion of spliceosomal snRNAs using a DNA oligonucleotide complementary to an accessible region of the snRNA and the extract endogenous RNaseH activity to degrade the RNA. Introduction of fluorescently labeled snRNA then allows for successful reconstitution activity. (b) smFRET investigation of U2 snRNA binding to the pre-mRNA in the presence of extract and ATP. (c) example traces showing short and long lived association of U2 snRNA with the pre-mRNA in the presence of extract and ATP.

into the $B^{act}$ complex for SiMPull-FRET as previously described. Such an experimental setup would allow for the direct observation of the Prp16 and ATP-dependent unwinding of the U2-BS duplex during alternative branchsite selection.

In addition to labeling of snRNA factors, the emergence of several new protein labeling techniques will allow for the observation of specific protein cofactors associating or dissociating with the spliceosome in response to progression through a particular splicing complex (**Table A.1**). By site-specifically labeling the pre-mRNA target or yeast snRNAs, binding proximity can be estimated as well as visualization of protein translocation across or through an RNA-RNA or RNA-protein duplex. Such experiments will be especially useful for investigating the mechanisms of action for the yeast ATPases, particularly Prp16 that is thought to possess multiple ATP-dependent and ATP-independent roles during and after the first step of splicing. One immediate application could be to label Cwc25 and monitor its Prp16-dependent stabilization with the spliceosome that is thought to occur in the presence of mutated branchsite sequences[41]. Other combinations of various pre-mRNA-snRNA, snRNA-snRNA, pre-mRNA-protein, and snRNA-protein labeling schemes, further combined with the specificity of SiMPull-FRET and computational power of SiMCAn, will allow for a bird's eye view of changes in RNA and protein conformation and thus reveal other mechanisms utilized by the spliceosome to achieve a high degree of specificity and efficiency during spliceosome assembly and catalysis. Furthermore, the use of more different colors in single-molecule experimentation will increase the amount of information extracted by, for example, using Pacific Biosciences SMRT Technology.

Implementation of the SHAPE-MaP approach to characterize the secondary structure of all yeast pre-mRNA transcripts within the $B^{act}$ and C complex will provide a nucleotide-

resolution picture of the pre-mRNA structure for nearly every yeast intron containing gene. Such knowledge will allow us to compare the structural changes of Ubc4 upon its profound remodeling from the distal $B^{act}$ to the proximal (and possibly more structure) C complex with that of all other actively spliced pre-mRNAs revealing any correlation between intron secondary structure and relative efficiency of splicing. Once established, this technique can in principle be applied to any purifiable RNA-protein complex, likely finding broad applications in the biomedical sciences.

Given the conservation of the splicing components between yeast and humans, the structure-function relationships and dynamics discovered in this thesis have the potential to also be utilized in humans and other higher eukaryotes. As protein and RNA labeling strategies become more accessible and broadly applicable, the number of potential targets for single molecule experiments will expand as will the complexity of the experiments and varieties of organisms capable of being studied, ultimately leading to a complete understanding of RNA structure and dynamics that can be properly and successfully targeted for therapeutic purposes.

# APPENDIX A: Identifying a mechanism of RNA unwinding by Prp22 in isolation and within the spliceosome using single-molecule FRET

## A.1 Introduction

Throughout the splicing cycle, numerous RNA-RNA and RNA-protein conformational rearrangements are required to ensure proper assembly, regulation, and catalysis. In the budding yeast *Saccharomyces cerevisiae,* these rearrangements are facilitated by a set of at least 8 known RNA helicases/ATPases of the DExD/H-box family of helicases that include Prp5, Prp28, Brr2, Prp2, Prp16, Prp18, Prp22, and Prp43[65]. Four of the known ATPases have been shown to possess *in vitro* helicase activity[92,193-195] and at least four are thought to allow for spliceosomal proofreading[42,116]. These proofreading events are thought to be one of the primary ways by which mutated substrates are detected and rejected and thus it is of great importance that we have a thorough understanding of how these rearrangements occur. The RNA helicases have been characterized primarily through observation of effects wild-type or mutant versions of the proteins have on yeast viability. Additionally, the *in vitro* helicase activities of Prp22 and Prp16 have been extensively characterized biochemically using model substrates. Despite all of this work, it remains unclear how the helicase of these enzymes are utilized in the spliceosome. We therefore have started to characterize the role of Prp22 in the late stages of the spliceosome by applying the tool of single-molecule fluorescence resonance energy transfer (smFRET).

Prp22 is a known RNA helicase responsible for mRNA release after the second step of splicing. It is thought that Prp22 uses its helicase activity to disrupt base-pairing between the

mRNA and the U5 snRNA. In order to first characterize mRNA-dependent unwinding and release from the spliceosome, we designed a short DNA probe that binds to the pre-mRNA Ubc4 to serve as a model substrate with which to study Prp22 helicase activity (**Error! Reference source not found.a**). By fluorescently labeling Prp22 and the DNA probe at its 3'end, we will be able to monitor Prp22 association with the duplex, ATP-dependent unwinding of the duplex by Prp22 (as seen through an increase in FRET), and the eventual complete dissociation of the labeled DNA oligonucleotide. In addition, by lowering the concentrations of ATP within the solution, the Prp22 helicase activity will be slowed, allowing for the observation of a stepwise increase in FRET as Prp22 unwinds the duplex. Such an approach has been successfully utilized to study a well-characterized DNA helicase T7 gp4[196]. These experiments will provide valuable information about the specific mechanism of Prp22-mediated unwinding using a model substrate with which we can then begin to use to study Prp22 activity in the context of the spliceosome (**Error! Reference source not found.b**).

A major hurdle with such an experiment is acquiring single, site-specific labeling of Prp22. Site-specific functionalization and labeling of proteins are central aims in protein engineering, allowing for the visualization of living systems in their native environments. Until recently, optical imaging of proteins in cells or cell extract has relied heavily on the utilization of fusion proteins such as GFP to successfully complete this task. Due to their large size (~230 amino acids), the use of fusion proteins to study the role of proteins in living systems is often difficult, giving the tendency for fusion proteins to interfere with protein folding, activity, and localization[197]. We therefore set out to investigate a variety of previously-established labeling approaches in order to identify the most universal technique capable of labeling a variety of splicing protein cofactors for use in single-molecule experimentation.

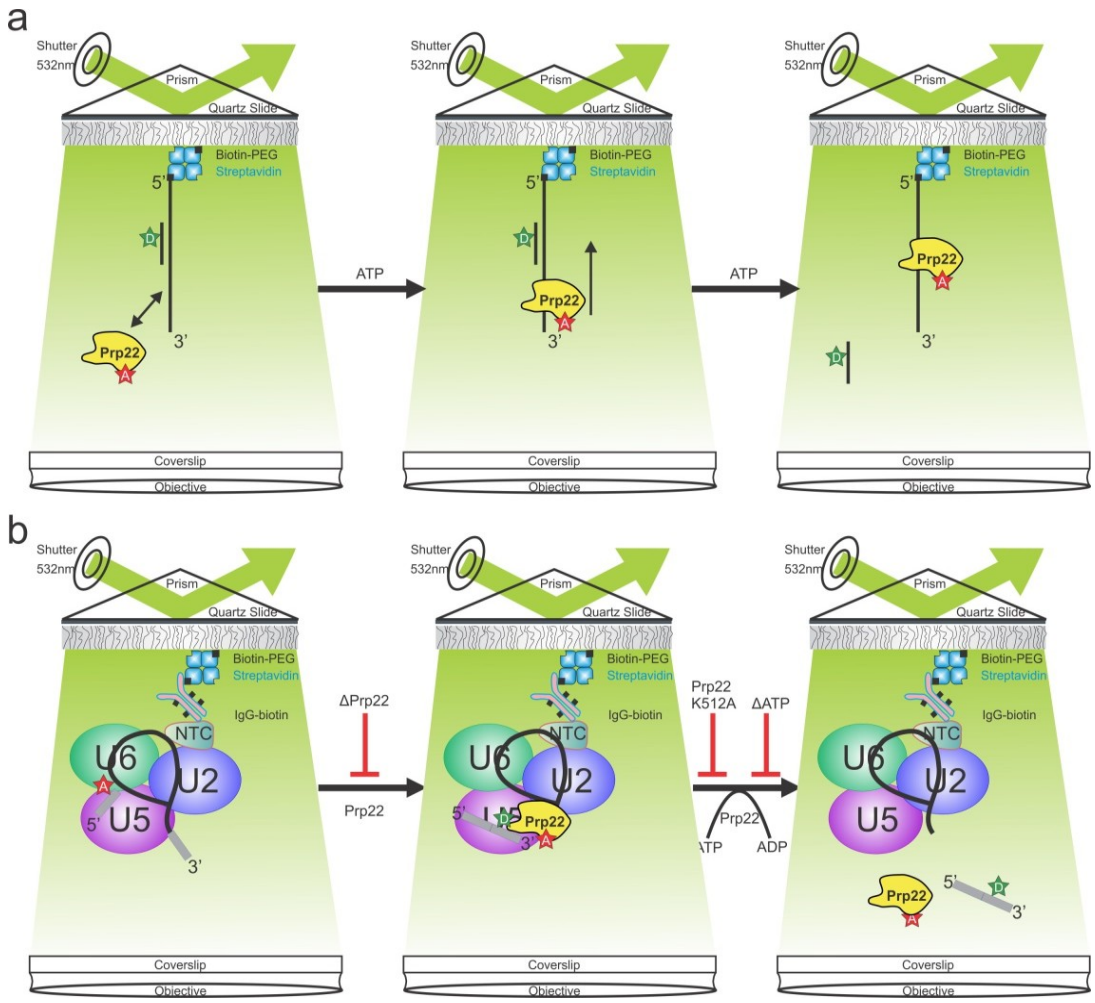



**Figure 6.2 Single molecule observation of Prp22-mediated unwinding using an optimized DNA-RNA hybrid and in the spliceosome**

Proposed single-molecule experiments to investigate Prp22's mechanism and function in unwinding a model duplex (a) and influencing mRNA release after both steps of splicing (b).

## A.2 Materials and methods

A.2.1 Expression and purification of intein-containing piece 1 and piece 2

Plasmids encoding the N-terminal piece 1 containing either Prp5, Prp2, Prp22, Cwc25, or Cus1 genes, as well as a 3' sequence encoding a hexahistidine tag, were produced through two rounds of PCR, restriction enzyme digestion, and ligation that first placed the gene of interest into the pRSETa plasmid and second the N-terminal portion of the SSP GyrB intein and hexahistidine tag. A plasmid encoding the C-terminal piece 2 containing the gene sequence for GB1 was made through PCR amplification of the GB1 gene using primers encoding the C-terminal portion of the intein and a single cysteine residue in the forward primer and a TEV cleavage site and hexahistidine tag in the reverse primer. The PCR product was cloned into the pRSETa plasmid for overexpression, purification and labeling (**Figure A.2**).

The five N-terminal piece 1 constructs and the C-terminal piece 2 protein were expressed and purified as previously described[92]. Briefly, cell pellets were resuspended in buffer A (50 mM Tris pH 7.5, 250 mM NaCl, 10% sucrose) and incubated with 0.2 mg/ml lysozyme with gentle stirring for 40 min. The suspension was adjusted to 0.1% Triton X-100 and subsequently removed of insoluble material through centrifugation at 18,000 rpm for 40 min. The supernatant was then incubated with Ni-NTA resin (Qiagen) with rotation for 1 h. The mixture was added to a column for removal of unbound material with repetitive washes with buffer E (50 mM Tris pH 7.5, 250 mM NaCl, 10% glycerol) containing 10 mM imidazole. Bound material was eluted with buffer E containing 20, 50, 100, and 500 mM imidazole. Peak protein fractions for Prp2, Prp22, and Prp5 were pooled and diluted with buffer D (50 mM Tris pH 7.5, 2 mM DTT, 1 mM EDTA, 10% glycerol) to adjust salt concentrations to 50 mM for purification on polyuridylic acid-agarose (Sigma). Resin was washed extensively with buffer D containing 50 mM NaCl and

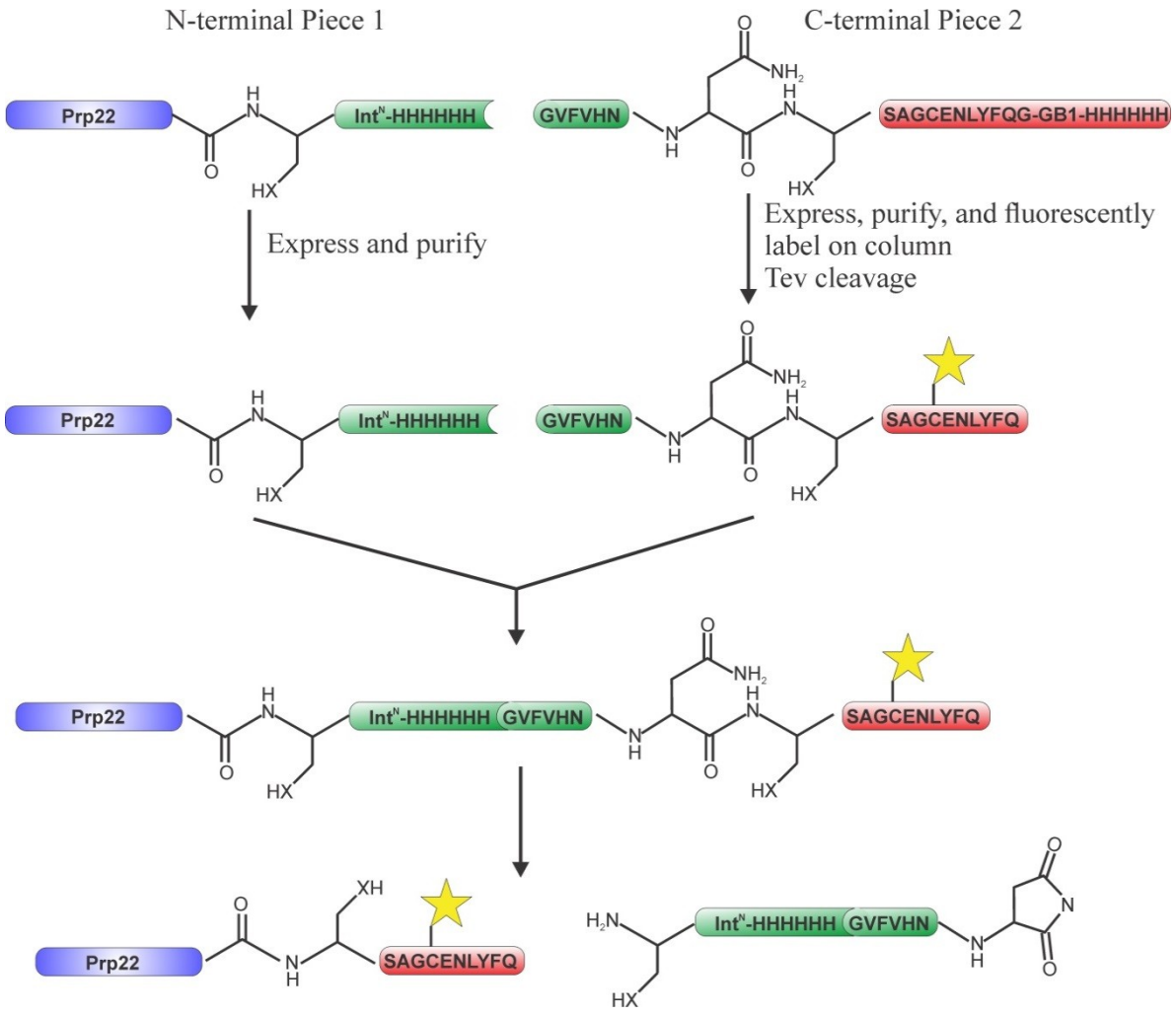



**Figure 6.3 Intein-mediated labeling strategy for Prp22**
Proposed single-molecule experiments to investigate Prp22's mechanism and function in unwinding a model duplex (a) and influencing mRNA release after both steps of splicing (b).

subsequently eluted with buffer D containing 100, 200, 300, and 500 mM NaCl. Finally, all proteins were dialyzed against buffer D containing 50 mM NaCl, concentrated, aliquoted, and flash frozen in liquid nitrogen. Protein concentrations were determined using the Bradford dye reagent (Bio-Rad).

The single-cysteine C-terminal piece two protein was labeled with Cy3-maleimide (GE Healthcare). Labeling was performed using 0.150 μmol of protein and 0.5 mg of dye containing 10 μM reducing agent Tris(2-carboxyethyl)phosphine (TCEP) (Sigma). Reactions were incubated at 23°C for 1 h followed by overnight at 4 ºC. Free dye was removed by re-purification of protein on a $Ni^{2+}$ column and dialysis back into buffer D. The extent of labeling was determined using A280 and A550 readings and found to be 90%.

A.2.2 Protein trans-splicing reactions

*In vitro* protein labeling reactions were performed as described[198]. Briefly, 1 uM N-terminal piece two protein containing Prp5 was incubated with 15 uM C-terminal piece two protein containing Cy3 in the optimized splicing buffer (oSB: 20 mM Tris-HCl pH 8.5, 250 mM NaCl, 1 mM EDTA) in the presence of 0.1 mM TCEP for 16 h at room temperature. Reactions were allowed to incubate with Ni-NTA resin for 1 h before washing with buffer E. Bound protein was then eluted using buffer E containing 150 mM imidazole. Wash and elution fractions were resolved on a 10% SDS-PAGE gel and analyzed by scanning on a Typhoon variable mode imager (GE Healthcare) followed by coomassie staining.

A.2.3 Prp22-Ald expression and purification

Prp22-Ald was expressed and purified essentially as described[92]. Briefly, cell pellets containing over-expressed Prp22-Ald were resuspended in buffer A (50 mM Tris pH 7.5, 250 mM NaCl,

10% sucrose) and incubated with 0.2 mg/ml lysozyme with gentle stirring for 40 min. The suspension was adjusted to 0.1% Triton X-100 and subsequently removed of insoluble material through centrifugation at 18,000 rpm for 40 min. The supernatant was then incubated with Ni-NTA resin (Qiagen) with rotation for 1 h. The mixture was added to a column for removal of unbound material with repetitive washes with buffer E (50 mM Tris pH 7.5, 250 mM NaCl, 10% glycerol) containing 10 mM imidazole. Bound material was eluted with buffer E containing 20, 50, 100, and 500 mM imidazole. Peak Prp22 fractions were pooled and diluted with buffer D (50 mM Tris pH 7.5, 2 mM DTT, 1 mM EDTA, 10% glycerol) to adjust salt concentrations to 50 mM for purification on polyuridylic acid-agarose (Sigma). Resin was washed extensively with buffer D containing 50 mM NaCl and subsequently eluted with buffer D containing 100, 200, 300, and 500 mM NaCl. Finally, Prp22 was dialyzed against buffer D containing 50 mM NaCl, aliquoted, and flash frozen in liquid nitrogen. Protein concentrations were determined using the Bradford dye reagent (Bio-Rad).

A.2.4 Prp22-Ald fluorescent labeling

Purified Prp22-Ald was fluorescently labeled with Cy5-hydrazide or CF640r as described[199]. Prp22-Ald was exchanged into labeling buffer (250 mM potassium phosphate, 500 mM KCl, and 5 mM DTT) using Amicon Ultra-0.5 centrifugal filter units (Millipore). Protein was then mixed with dried fluorescent dye and incubated at 4 °C for 18 h. Unincorporated dye was removed using either Micro Bio-spin P-6 or P-30 columns (Bio-Rad), Centri-Spin 10 columns (Princeton Separations) or by re-purifying over Ni-NTA in the absence or presence of 2 M urea. Extent of labeling and dye removal was determined through analysis on 10% SDS-PAGE.

## A.3 Results

### A.3.1 Intein-mediated protein labeling

In recent years, several approaches have utilized the attachment of 'reporter handles' to target proteins followed by the modification with exogenously added probes (**Table A.1**)[200]. To qualify as a sufficient labeling method, these reporter handles should not perturb the folding and activity of the protein, require a high degree of specificity, be relatively small, undergo rapid highly chemoselective reactions, and should be able to occur in physiological conditions. One such approach is protein trans-splicing (PTS), also known as intein-mediated protein ligation. Protein splicing is a naturally occurring process in which a protein editor, called an intein, excises itself out of a host protein in which it is embedded creating a new peptide bond between its two flanking regions, the exteins[201]. In the 20 years since its discovery, protein splicing has been utilized for the development of several protein-engineering methods, one of these being protein trans-splicing (PTS). PTS uses an artificially or naturally split intein to create a new peptide bond between flanking exteins[202]. Split inteins are characterized by the fact that their primary amino acid sequence is cut into two polypeptides, an N-terminal fragment and a C-terminal fragment (**Figure A.2**). Unlike other protein splicing techniques such as expressed protein ligation (EPL) and native chemical ligation (NCL), PTS does not require extremely high concentrations of reactants, can occur in native conditions, and does not require a thioester or N-terminal cysteine on the target protein[201]. Additionally, inteins split very near the N- or C-terminus can utilize short, synthetic peptides as one component in the reaction allowing for the incorporation of many chemical modification groups into the synthesized peptide. One such split intein is the SSP GyrB mini-intein (DNA gyrase subunit B from Synechocystis species)[203]. This intein is known to contain a natural cleavage site, but has also been artificially split six amino acids in from the C-

| Label Type | Pros | Cons |
|---|---|---|
| **Snap Tag** | • Rate of labeling is independent of the size of the attached unit<br>• Can efficiently work with *in vitro* experiments<br>• Covalently labels a protein<br>• Fast reacting in bacteria and yeast<br>• Commercially available<br>• Can attach various different fluorophores without having to alter the protein | • Large protein (22kDa, 182 aa) |
| **Formylglycine Generating Enzyme** | • Has been shown to work with a small, six amino acid tag<br>• Aldehyde can be reacted with a variety of hydrazide-functionalized fluorophores<br>• Highly efficient (90-99%) | • Potential harsh labeling conditions<br>• Requires high concentrations of dye |
| **Intein-mediated Protein Ligation** | • The IMPACT kit is commercially available from NEB<br>• Label N- or C-terminus using a small peptide<br>• Fast and efficient labeling<br>• Small (~10 amino acids) labeling tag | • Might be protein specific<br>• Protein solubility problems<br>• Requires high protein concentrations |
| **Unnatural Amino Acid Incorporation/Amber Codon suppression mutagenesis** | • Very selective<br>• Minimal perturbation to protein structure<br>• Versatility in terms of the small molecule label that can be used<br>• Can potentially label any part of the protein | • Must generate the tRNA/synthetase pair<br>• Natural prevalence of amber codons in eukaryotes |
| **Flash and Cy3AsH** | • Cell-permeable<br>• Dissociation constants are sub-nanomolar<br>• N or C-terminus incorporation<br>• Commercially available (Flash) | • Requires additional reagents to reduce background<br>• Potential cellular toxicity<br>• Limited photostability and very pH sensitive |
| **Halo Tag Mediated Labeling** | • Neither E. coli nor eukaryotic cells have endogenous dehalogenases<br>• HaloTag ligands can be easily labeled with small organic dyes | • Largest of all covalent attachments (33 kDa) |
| **Biotin Ligase** | • Natural proteins don't have keto groups.<br>• Tag is small (15 amino acids)<br>• Ease of synthesis of hydrazide derivatives<br>• Good for proteins that are affected by recombination with other proteins | • Second step of labeling VERY slow at pH 7<br>• Not cell permeable<br>• Very low yield for the synthesis of ketone-1 used instead of biotin |
| **Lipoic Acid Ligase** | • Allows for covalent attachment<br>• Highly specific and fast reacting<br>• Does not require a large protein target<br>• Peptide target (22 amino acids) is recognized at N- or C-terminus<br>• Alkyl azides are resistant to oxidation and don't cross-react with amines. | • Potentially poor yield<br>• Protein end-labeling only<br>• Peptide Sequence recognized is 22 amino acids<br>• Intracellular components can not yet be labeled<br>• A two-step labeling process |

**Table 6.1 Protein labeling approaches**
A list of pros and cons for several protein labeling strategies developed over the last 10-20 years. Although it has been shown to possibly be the fastest and efficient form of labeling, the SNAP tag was not pursued due to its large size.

terminal end of the protein and shown to efficiently reconstitute protein splicing upon assemblage of the two subunits[198,204].

In order to adapt the SSP GyrB split intein for labeling of splicing factors, the Prp22 gene of interest was cloned into an expression plasmid containing a region encoding the N-terminal portion of the SSP GyrB intein as well as a C-terminal hexahistidine tag known not to affect intein recognition and splicing (**Figure A.2**, N-terminal Piece 1). Furthermore, an additional plasmid was developed that encoded for the C-terminal portion of the intein (GVFVHN) followed by a unique cysteine residue, a TEV cleavage site (ENLYFQG), the GB1 stabilizing protein, and a hexahistidine tag (**Figure A.2**, C-terminal Piece 2). Upon expression and purification of both constructs, as well as labeling and TEV cleavage of the C-terminal intein piece 2, we anticipated that PTS would result in formation of a labeled splicing factor Prp22 lacking a hexahistidine tag. Spliced, labeled product could then be purified away from unreacted starting material through a final purification on a $Ni^{+2}$ column and passage through a gel filtration column. Similar constructs were developed encoding additional splicing factors including Prp2, Prp5, Cus1, and Cwc25. Unfortunately, it was found that many protein conjugates containing the N-terminal portion of the intein are very unstable, thus preventing purification and storage of highly concentrated protein required for PTS. Fortunately, Prp5-containing N-terminal piece one was found to be stable enough at high concentrations for initial PTS trials using a Cy3-labeled C-terminal piece two. To initiate PTS, the two halves of the intein were combined with a 10-fold excess of labeled piece two and allowed to react for 18 hours at room temperature in the presence of 0.1 mM TCEP[198]. The splicing reaction mixture was then incubated with Ni-NTA resin to allow the unreacted Prp5 starting material to bind the column. Interestingly, washing the column with a mild buffer resulted in release of a significant amount

of protein around the expected molecular weight of 100 kDa that was also visible in a Cy3 scan of the gel (**Figure A.3,** Prp5-Cy3 band). Eluting bound material with high concentrations of imidazole resulted in release of a band around 120 kDa corresponding to the Prp5 starting material that was visible in the coomassie stain but not the Cy3 scan. This result would appear to indicate that the two pieces of the intein successfully reacted at low levels to release labeled Prp5 lacking a hexahistidine tag that was only detectable in the wash fractions of the column; unlabeled starting material remained bound to the column and eluted only upon addition of imidazole.

Although this approach does appear to work for the labeling of Prp5, intein-mediated protein labeling does not appear to be as universal of a labeling approach as needed to label other important yeast splicing factors such as Prp22 and Prp16. Future work will require the labeled, C-terminal piece 2 to be more efficiently purified to remove the many impurities that become more easily visible upon labeling with a fluorescent dye (**Figure A.3**). As a result, the PTS reaction and the downstream purification from unreacted C-terminal piece two using gel filtration will become much more efficient.

A.3.2 Utilizing Formylglycine Generating Enzyme to Fluorescently Label Prp22

Another promising protein labeling method is the use of the formylglycine generating enzyme (FGE)[205]. FGE is an enzyme that catalyzes the oxidation of a cysteine residue in a 6-12 amino acid sequence[206] of sulfatases to a formylglycine residue to form the mature active site of the sulfatase[207]. This recognition sequence, termed the 'aldehyde tag,' has been found to still be recognized and modified by FGE when cloned onto the ends or internal regions of various proteins (**Figure A.4a**)[208,209]. Additionally, the FGE labeling approach has recently been utilized for the labeling of a DNA modification enzyme for use in single-molecule studies[199]. Thus, we

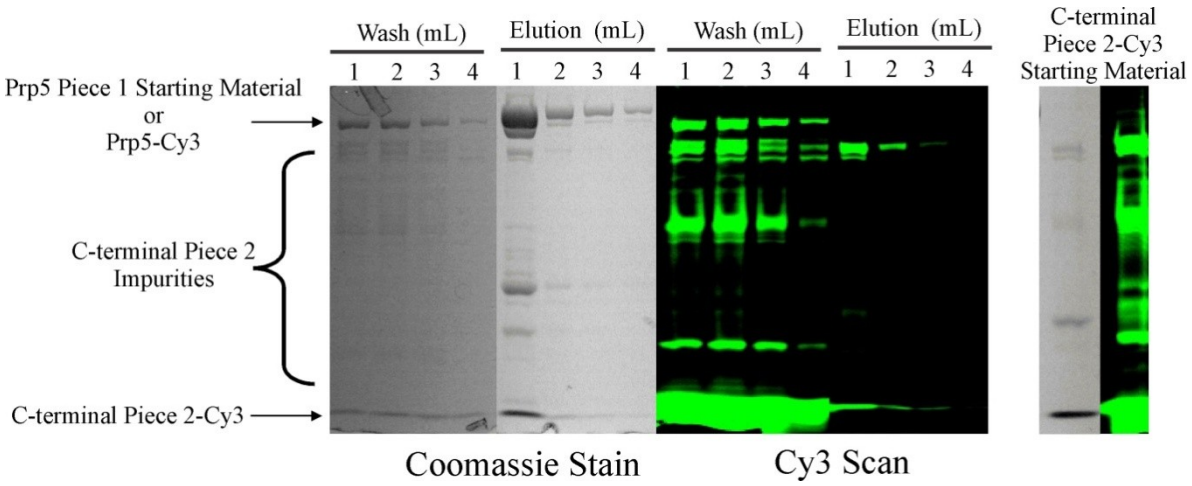



**Figure 6.4 Spliced and labeled Prp5 elutes in Ni-NTA wash fractions**
PTS reactions between Prp5-containing N-terminal Piece 1 and Cy3-labeled C-terminal Piece 2 were loaded onto a Ni-NTA column for purification of spliced material. Spliced, labeled Prp5 (Prp5-Cy3, 100 kDa) came off the column in the wash fractions while the slightly larger starting material (Prp5 Piece 1, 120 kDa) remained on the column until elution with imidazole. The many impurities, primarily visible in the Cy3 scan, can be traced back to the C-terminal Piece 2-Cy3 starting material.

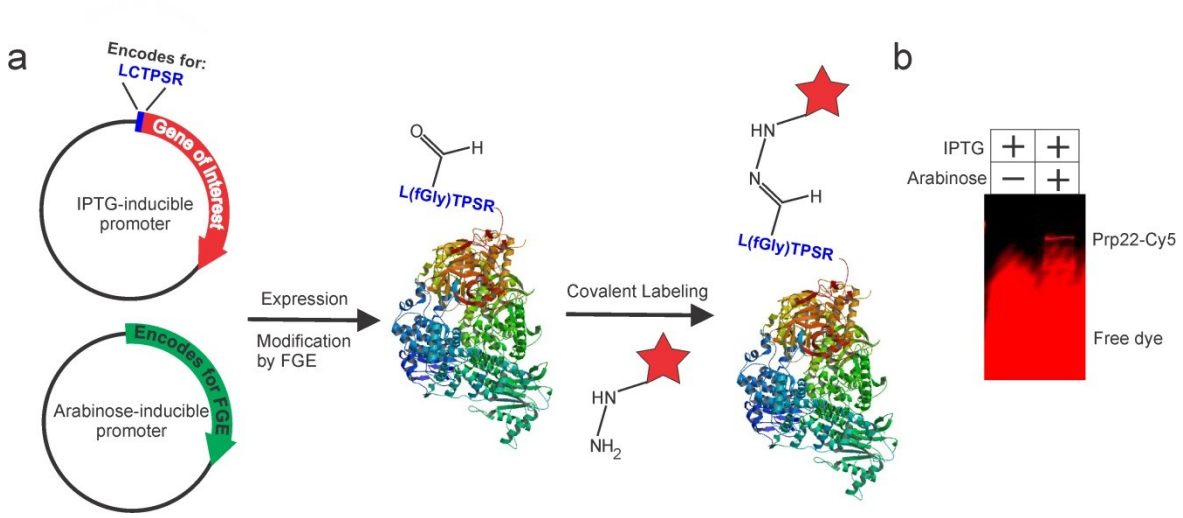



**Figure 6.5 FGE-mediated fluorescent labeling strategy for Prp22**
(a) Co-expression of FGE and an aldehyde tagged protein of interest results in modification of the active site cysteine residue to formylglycine. Incubation of the modified protein with hydrazide-derivative fluorophores results in site-specific, fluorescent labeling. Shown is the Prp8 crystal structure as an example of a splicing protein to label (PDB accession number 4I43). (b) Labeling of Prp22-Ald with Cy5-hydrazide dye is specific to Prp22 only when it is co-expressed with the modifying protein FGE.

speculated that the 'aldehyde tag' could be a valuable tool through which to modify and label spliceosomal proteins in isolation and potentially in yeast cell extract as described for the SNAP-tag[110].

To adopt the FGE-mediated labeling strategy for labeling of Prp22, we cloned a DNA sequence encoding the minimal six amino acid aldehyde tag onto the 3' end of Prp22 along with a hexahistidine tag and placed the sequence into a ampicillin resistant plasmid containing an IPTG-inducible promoter for overexpression and purification in bacteria (from here on known as Prp22-Ald). Additionally, the gene sequence encoding the FGE protein was cloned into a kanamycin resistant, arabinose-inducible plasmid so that both plasmids could be selected for using the appropriate antibiotics and selectively-expressed using the desired inducing agents (**Figure A.4a**). Cell extracts were made from bacteria that were either incubated with just IPTG during growth or IPTG and arabinose to allow expression of Prp22 in the absence and presence or FGE. Interestingly, incubation of cell extracts with Cy5-hydrazide dyes resulted in fluorescent labeling of Prp22 only when FGE was co-expressed with Prp22 (**Figure A.4b**). The –FGE lane was completely absent of labeled protein and contained only unincorporated dye. These data demonstrate the specificity of the labeling strategy and also indicate that labeling of Prp22 might possibly be achieved in yeast splicing extract after modification by FGE.

The proposed single-molecule experiments (**Figure A.1**) require a purified, labeled protein void of any traces of remaining free dye. We therefore co-overexpressed and purified Prp22-Ald for labeling with Cy5-hydrazide. Recent optimization of the aldehyde tag labeling strategy identified a more neutral pH buffer (250 mM potassium phosphate, 500 mM KCl, and 5 mM DTT)[199] than what was used in initial testing (100 uM MES pH 5.5, 1% SDS)[208]. Purified Prp22-Ald was buffer exchanged into the optimized labeling buffer and incubated with Cy5-

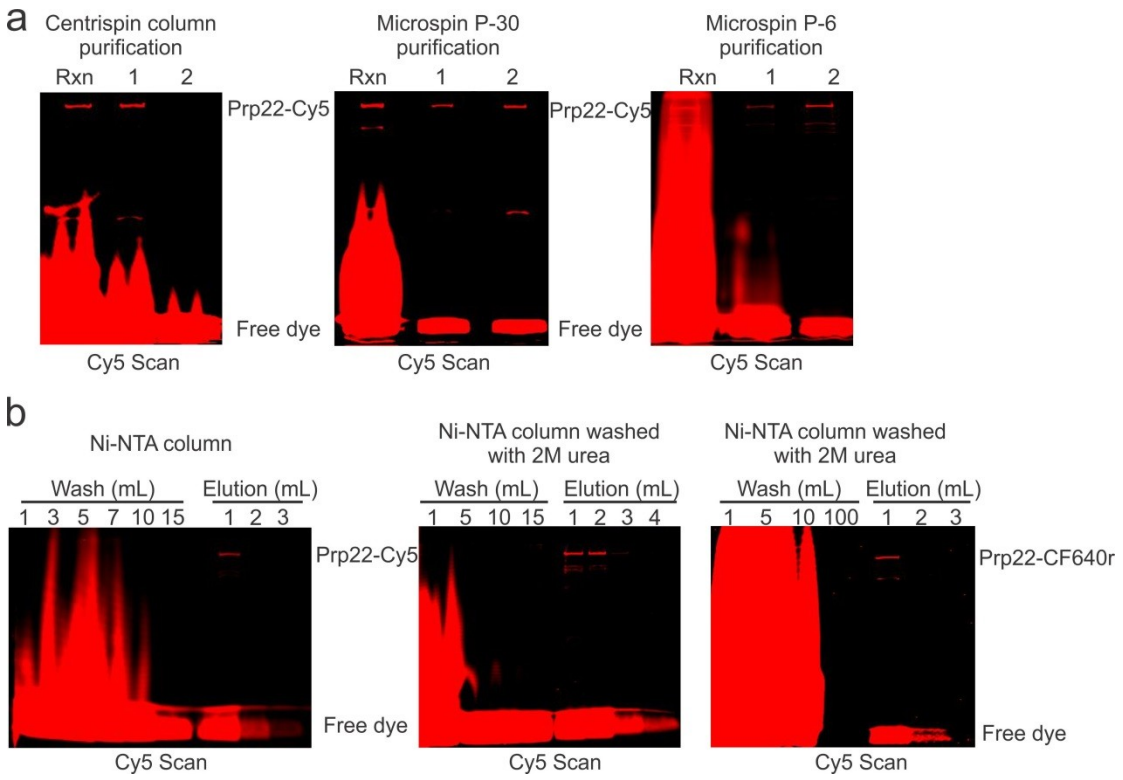



**Figure 6.6 Ni-NTA and gel filtration columns are not sufficient for removing free dye**

hydrazide for 12 hours. We then attempted to remove the large excess of free dye through use of a number of gel filtration spin columns as previously reported (**Figure A.5a**) as well as through re-purification on Ni-NTA resin (**Figure A.5b**). Unfortunately, even after passing through multiple rounds of spin columns and after more than 15 mL of washing with a high-salt wash buffer, a significant amount of free dye remained stuck to the protein (**Figure A.5a,b)**. Believing the cause of the remaining free dye to be through hydrophobic sticking of the dye to the protein, 2 M urea was included in the wash buffer during purification on Ni-NTA resin. Interestingly, washing with over 15 mL of the strong denaturant still did not remove the remaining free dye. Lastly, we attempted labeling with CF640r, a much less hydrophobic dye than Cy5-hydrazide but with similar spectroscopic properties[210]. Surprisingly, use of the CF640r fluorophore and purification on Ni-NTA with 2 M urea still resulted in free dye sticking to Prp22 and coming out in the elution fractions (**Figure A.5b**, far right).

### A.4 Discussion

Site-specific functionalization and labeling of proteins are central aims in protein engineering as well as single-molecule investigation of protein function in complex RNA-protein systems. Recently, several single-molecule groups have utilized the high specificity and catalytic efficiency of SNAP-tag-mediated protein labeling to monitor assembly of specific snRNP complexes[110]. This method has a great advantage in that it is very rapid and can also take place in yeast whole cell extract allowing for the incorporation of SNAP tags onto a variety of splicing protein cofactors directly in the yeast genome prior to purification of extract. Unfortunately, the large size of the SNAP tag may perturb protein function and, as a consequence, splicing assembly or catalysis will be defective. We thus set out to experiment a number of protein labeling strategies that result in the attachment of a small reporter handle (~6 amino acids)

containing a variety of fluorescent dyes to several essential spliceosomal proteins. The first of these was protein trans-splicing (PTS) using a split intein. The SSP GyrB intein was recently split six amino acids upstream from the C-terminus and found to still efficiently reconstitute PTS upon mixing of the two portions of the intein. The small size (six amino acids) of the C-terminal portion of the intein allows for synthesis of small peptides containing the C-terminal sequence with a variety of attached modifications (i.e., fluorophores). Unfortunately, peptide synthesis with Cy3 and Cy5 fluorophores, those most commonly used for single-molecule FRET, is difficult, expensive, and often results in production of an insoluble peptide. Therefore, we developed a plasmid encoding the C-terminal portion of the intein immediately upstream of a cysteine residue, a hexahistidine tag, and GB1, a small, stable, single domain protein that, most importantly, lacks cysteine residues. Expression of this construct allowed for the purification and labeling of large amounts of the protein containing a variety of fluorescent probes at a relatively low cost. Our strategy was also designed to easily separate spliced from unspliced material as a result of PTS. Unfortunately, the high concentrations (uM) of starting material became an issue for a majority of the spliceosomal proteins attempted, making the approach not as universal as needed to effectively study the protein components of the spliceosome by smFRET.

Second, we attempted labeling of Prp22 containing a six amino acid recognition sequence for the formylglycine generating enzyme (FGE). FGE binds and modifies a cysteine residue within the recognition sequence to a unique aldehyde that can subsequently be labeled with hydrazide-containing fluorophores. This labeling strategy was recently utilized to fluorescently label a DNA binding protein for analysis by smFRET[199]. The authors also revealed an optimized labeling buffer that allows for efficient labeling at neutral pH. Intriguingly, we were able to achieve efficient labeling of Prp22 carrying the short ‘aldehyde tag’ under the conditions

described. However, removal of free dye from such a large protein (~130 kDa) proved to be very difficult and impeded our ability to use labeled Prp22 in single-molecule experiments. Several fluorescent dyes and gel filtration strategies were attempted to remove unincorporated dye with little success. As a result, fluorescent labeling of large, and often hydrophobic, proteins may prove to be very challenging regardless of the labeling strategy.

An additional strategy we propose to utilize in the future is the use of nonsense suppression to incorporate unnatural amino acids (UAA) containing a variety of unique functional handles. Such a strategy has been successfully used for the site-specific incorporation of BPA and azide functional groups into yeast proteins[211-213]. Several fluorescent dyes containing the corresponding alkyne functional groups are commercially available that could allow for the labeling of spliceosomal proteins in yeast splicing extract or in an isolated solution of purified protein.

# APPENDIX B: Observing Prp28-dependent Changes in pre-mRNA Conformation in Early Spliceosome Formation[5]

## B.1 Introduction

The spliceosome is the megadalton protein-RNA complex responsible for the complex removal of nearly all noncoding RNA segments of precursor messenger RNA (pre-mRNA) transcripts. Spliceosome assembly is thought to take place in a stepwise manner in which the five small nuclear ribonucleoprotein (snRNP) complexes assemble and reorganize to form a complex network of RNA-RNA and RNA-protein interactions that comprise a catalytic core capable of carrying out both chemical steps of splicing. Central to the assembly process are the eight DExD/H-box ATPases responsible for catalyzing, in an ATP-dependent manner, several conformational rearrangement steps that are also thought to be points of spliceosomal proofreading. Interestingly, recent work has revealed that several of the conserved spliceosomal ATPases have additional ATP-independent roles vital to the assembly and catalytic processes throughout the splicing cycle. One recent example of this dual-role nature is Prp28, a DEAD-box ATPase that promotes U1 snRNP release prior to $B^{act}$ complex formation. Prp28 was recently found to play a minor role during the ATP-independent formation of the commitment complexes[214]. The commitment complexes (CC1 and CC2) were originally identified as the earliest-forming stable complexes that commit the pre-mRNA to splicing and are resistant to

[5] Matthew Kahlscheuer and Ramya Krishnan performed the smFRET experiments on the Ubc4 pre-mRNA. Argenta Price prepared all yeast splicing extracts, Prp28 protein, and performed the native gel analysis of the Ubc4 pre-mRNA substrate.

competition from addition of naked pre-mRNA[215]. These complexes form in the absence of ATP and contain either just the U1 snRNP bound to the 5'SS (CC1) or additionally contain the branchpoint binding protein (BBP) and Mud2 bound to the branchsite sequence (CC2). Interestingly, while depletion of BBP/Mud2 from extract does prevent CC2 formation, complete assembly, catalytic activation, and both steps of splicing can proceed unimpeded in the presence of ATP[216]. CC2 formation may, therefore, not be strictly required for splicing under optimal splicing conditions. Furthermore, recent native gel analysis of the commitment complexes revealed a dramatic loss of CC2 formation upon depletion of Prp28 from yeast splicing extract depleted of ATP while wildtype extracts depleted of ATP efficiently form CC2[214]. Unfortunately, native gel analysis may facilitate the stabilization of the commitment complexes with gel-caging interactions and thus do not allow for a thorough investigation of commitment complex stability or spliceosome and pre-mRNA conformation, preventing the determination of a specific role for Prp28 in CC2 formation.

Single-molecule fluorescence resonance energy transfer (smFRET) is a powerful biophysical tool that has recently been utilized to study and monitor pre-mRNA conformation and dynamics throughout spliceosome assembly and catalysis[53,55,62]. Through labeling of the Ubc4 BS and 5'SS with donor (Cy3) and acceptor (Cy5) fluorophores, respectively, time- and ATP-dependent changes in proximity of the points of first step chemistry can be observed. Given the highly time-sensitive nature of smFRET to changes in donor-acceptor distance, smFRET is an ideal tool with which to identify a Prp28-dependent role in altering pre-mRNA structure during CC2 formation. We therefore set out to investigate a potential role for Prp28 remodeling of the pre-mRNA substrate during formation of CC2. Utilizing a Prp28-depleted extract and the parent, wildtype extract, we initially observed a dramatic shift in FRET conformation from a

primarily low FRET state in the absence of Prp28 (CC1) to a dominant high FRET state upon introduction of the wildtype extract (CC2). We further investigated this shift in FRET utilizing a BP mutant substrate incapable of forming CC2 and observed a similar, low FRET conformation. However, upon re-analysis of the smFRET data and upon extensive experimentation with a variety of Prp28 reconstitution conditions, we confirmed that there appears to be no significant change in pre-mRNA conformation during assembly of CC1 and CC2. In addition, the presence or absence of Prp28 does not appear to have a significant effect on pre-mRNA conformation early in spliceosome assembly.

## B.2 Materials and Methods

### B.2.1 Preparation of fluorescently labeled pre-mRNA substrates

The Ubc4 pre-mRNA substrates used in this study (**Table B.1**) were synthesized as previously described[53]. Briefly, the 135-nucleotide pre-mRNA was ligated from two fragments: a 59-nucleotide 3' segment with 5-amino-allyl-uridine at the +6 position relative to the BP adenosine and a 76-nucleotide 5' segment with 5-amino-allyl-uridine at the -7 position relative to the 5'SS. The BP mutant had the branchsite adenosine at position 89 on the 3' segment replaced with cytosine. The 5' and 3' fragments were coupled to Cy5 and Cy3 N-hydroxysuccinimidyl ester (GE Healthcare), respectively, by resuspending 4 nanomoles of RNA in 40 µl of 0.1 M sodium bicarbonate buffer, pH 9.0, and incubating for 30 min at 60 °C with the proper dye pack dissolved in DMSO. The conjugated fragments were ethanol precipitated and washed with 70% (v/v) ethanol to remove unconjugated dye. Unlabeled RNA was removed by purification on benzoylated naphthoylated DEAE (BND)-cellulose (Sigma) that was washed with 1 M NaCl containing 5% (v/v) ethanol. Fully labeled RNA fragments were eluted with 1.5 M NaCl containing 20% (v/v) ethanol and further precipitated to remove excess salt. Labeled fragments

| | |
|---|---|
| Ubc4 Wildtype (WT) | 5'-biotin-GAACUAAGUGAUC**(5-N-U)**AGAAAG*GUAUGUCUAAAGU UAUGGCCACGUUUCAAAUGCGUGCUUUUUUUUUAAAACU UAUGCUCUUAUUUACUA* ***A*** *CAAAA* ***(5-N-U)*** *CAACAUGCUAUUG AACUAG* AGAUCCACCUACUUCAUGUU-3' |
| Ubc4 BP mutant | 5'-biotin-GAACUAAGUGAUC**(5-N-U)**AGAAAG*GUAUGUCUAAAGU UAUGGCCACGUUUCAAAUGCGUGCUUUUUUUUUAAAACU UAUGCUCUUAUUUACUA* ***C*** *CAAAA* ***(5-N-U)*** *CAACAUGCUAUUG AACUAGA* GAUCCACCUACUUCAUGUU-3' |
| DNA splint | 5'-GTTGATTTTGTTAGTAAATAAG(SP9)GTTTTAAAAAAAAAGCACGC-3' |

**Table 6.2 Sequence information of the oligonucleotides used in this study**
The Ubc4 intron is italicized, and the BP adenosine is bold and underlined. The red and green "(5-N-U)" denote the allyl-amine modified uridines used to attach the Cy5 and Cy3 fluorophores. In the BP mutant, the bold and underlined cytosine replaces adenosine in the wild-type sequence. The DNA splint is the oligonucleotide used for templated ligation during synthesis of the WT and 3'SS pre-mRNA substrates. Sp9 denotes a 9-carbon linker.

were combined with an equal molar amount of DNA splint (**Table B.1**) and ligated by incubating with RNA Ligase 1 (NEB) for 4 h at 37 °C as described[53,117]. Full length, labeled Ubc4 was then purified on a denaturing 7 M urea, 15% (w/v) polyacrylamide gel.

B.2.2 Preparation of yeast splicing extract and Prp28 protein

Splicing active whole cell extract (WCE) and Prp28 protein was prepared as described[214]. Wildtype extract was from Gal-driven Prp28 yeast (yPR88) grown in YEP+galactose (and shifted to fresh galactose for the final 3-5 hours). Prp28 depleted extract was the same strain shifted to YEP+glucose for 3-5 hours. Cells were then harvested and washed in AGK buffer (10 mM HEPES-KOH, pH 7.9, 1.5 mM $MgCl_2$, 200 mM KCl, 10% (v/v) glycerol, 0.5 mM DTT, 0.6 mM PMSF, and 1.5 mM benzamidine). A thick slurry of cells was dripped into liquid nitrogen to form small cell pellets that could be stored at -80 °C. The frozen pellets were disrupted using a ball mill. The resulting frozen powder was thawed in an ice bath and centrifuged at 17,000 rpm in a type 45 Ti Beckman rotor. The supernatant was then centrifuged at 37,000 rpm in a Ti-70 rotor for 1 h. The clear middle layer was removed with a syringe and dialyzed for 4 h against 20 mM HEPES-KOH, pH 7.9, 0.2 mM EDTA, 0.5 mM DTT, 50 mM KCL, 20% (v/v) glycerol, 0.1 mM PMSF, and 0.25 mM benzamadine with one buffer exchange.

B.2.3 Native gel analysis of early commitment complex formation

Commitment complex gels were run as previously described[214] except the pre-mRNA used was P32 labeled Ubc4 pre-mRNA. Both wildtype and branchsite mutant substrates were transcribed with P32 UTP using the Ambion Megascript T7 kit. ATP was depleted from extract by treatment with Hexokinase and 2 mM glucose (Fisher) followed by a second round of dialysis before extracts were aliquoted and frozen.

B.2.4 Single-molecule FRET experiment

Single Molecule FRET was carried out in the same manner as previously described[53,55]. Using a prism-based TIRF microscope[45,49,122], we collected data from single molecules incubated with Prp28-depleted or wildtype extract. Data were collected from five to seven fields of view 15 minutes after addition of extract. The donor (Cy3) and acceptor (Cy5) fluorophores were excited using a 532- and 635-nm laser, respectively, with the resulting emission recorded at 100 ms time resolution with a Princeton Instruments, I-PentaMAX intensified CCD camera. A FRET value was calculated by dividing the intensity of the acceptor emission by the total emission from both donor and acceptor.

## B.3 Results

B.3.1 P32-labeled Ubc4 recapitulates previous commitment complex formation results

The previous finding of a Prp28-dependent role in CC2 formation was performed using the RP51 yeast pre-mRNA substrate[214]. Before performing the smFRET analysis of CC formation, we wanted to test whether Ubc4 behaves in a similar manner or if this result is substrate specific. Therefore, we *in vitro* transcribed and 32P labeled the modified Ubc4 substrate sequence for analysis in CC gels. Perhaps not surprisingly, the Ubc4 pre-mRNA behaved in a similar manner to the RP51 substrate (**Figure B.1**). Depletion of ATP (wt-Gal –ATP lane) results in about a 50:50 mixture of CC1 and CC2 whereas depletion of Prp28 (28D –ATP lane) severely inhibits CC2 formation. Furthermore, the branchsite mutant Ubc4 substrate was found to only form CC1 in both ATP-depleted conditions (wt-Gal –ATP BrC) and Prp28-depleted conditions (28D –ATP BrC). These results show that BrC pre-mRNA and Prp28-depleted extracts result in reliable formation of a CC1-like complex while addition of ATP-depleted extract to a wildtype substrate allows formation of an equal amount of CC1 and CC2.

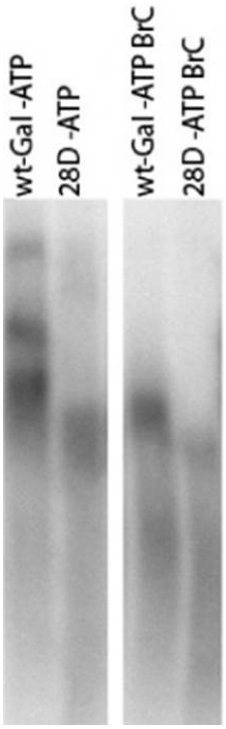



**Figure 6.7 Native gel analysis of commitment complexes using WT and branchsite mutant Ubc4**
WT or branchpoint mutant (BrC) 32P-labeled Ubc4 pre-mRNA was incubated with the indicated extract condition and resolved on native gels for identification of commitment complex formation.

B.3.2 Reconstitution of ΔPrp28 extract leads to a high FRET conformation

To first identify a Prp28-dependent role in pre-mRNA remodeling upon CC2 formation, we performed smFRET experiments with either a WT or BP mutant Ubc4 substrate fluorescently labeled near the BS and 5'SS with the FRET pair Cy3 and Cy5, respectively. Our prediction, based on the cross-intron bridging interactions between BBP and the U1 factor Prp40[217], was that the 5'SS and branch site should be brought into closer proximity in CC2. Substrates were immobilized on the slide surface and incubated with extract depleted of Prp28 (ΔPrp28-WCE (WT,BP)) and either reconstituted with recombinant Prp28 (+rPrp28) or with buffer (-Prp28) (**Figure B.3a,b,c**). Addition of Prp28 would be expected to result in CC2 formation with WT substrates but not with the BP mutant substrate which might result in a change in pre-mRNA conformation. Interestingly, these preliminary experiments revealed a dramatic shift to a dominant high FRET conformation in the presence of Prp28 with the WT substrate but not with the BP substrate (**Figure B.3a,b**). Extract depleted of Prp28 were characterized by a 50:50 distribution of a low, 0.10 and a high, 0.70 FRET state. These initial data are consistent with our model that Prp28 would promote formation of a higher-FRET CC2.

B.3.3 Repeat Prp28 reconstitution experiments were not able to reproduce our initial findings

Given our promising initial findings, we proceeded to test a variety of Prp28 reconstitution conditions and concentrations to confirm the shift in FRET state was a Prp28-dependent affect. Again, immobilized wildtype Ubc4 pre-mRNA substrates were incubated with ΔPrp28 extracts in the presence of a titration of increasing Prp28 concentrations. Surprisingly, all reconstitution conditions appeared to result in a predominant high FRET conformation as did reconstitution with buffer alone (**Figure B.3a**). Addition of low and high concentrations of Prp28 had very little effect on the distribution of FRET states or the associated dynamics as determined by

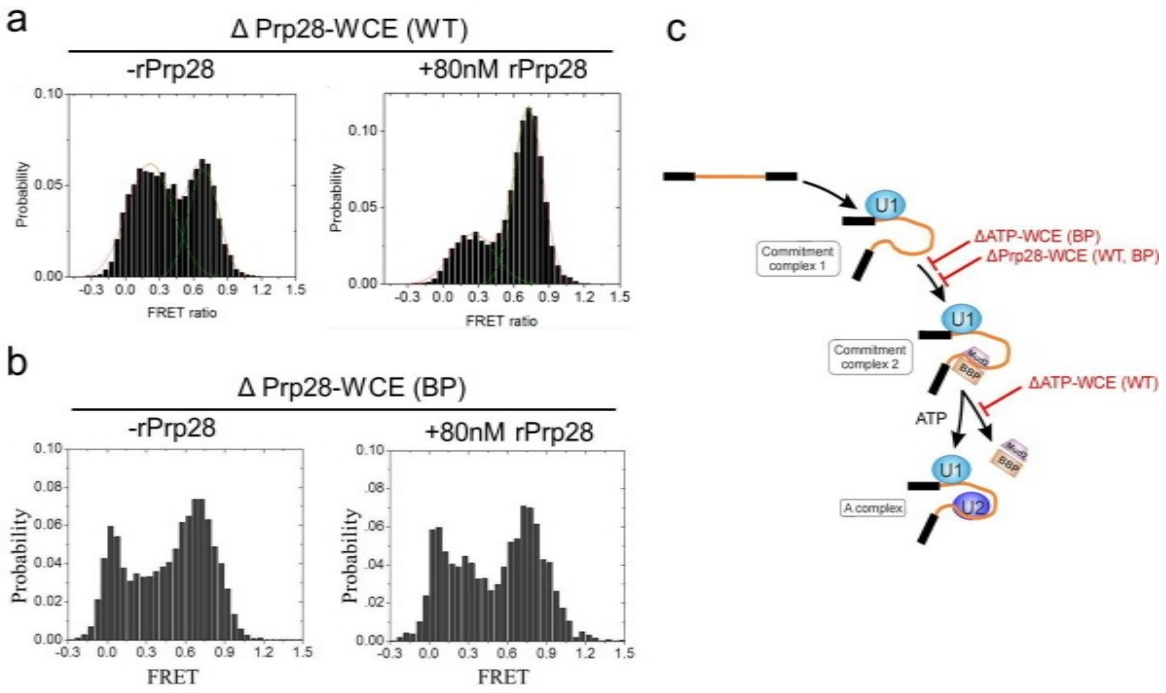



**Figure 6.8 Reconstitution with Prp28 appears to result in a high FRET conformation on WT pre-mRNA substrates**

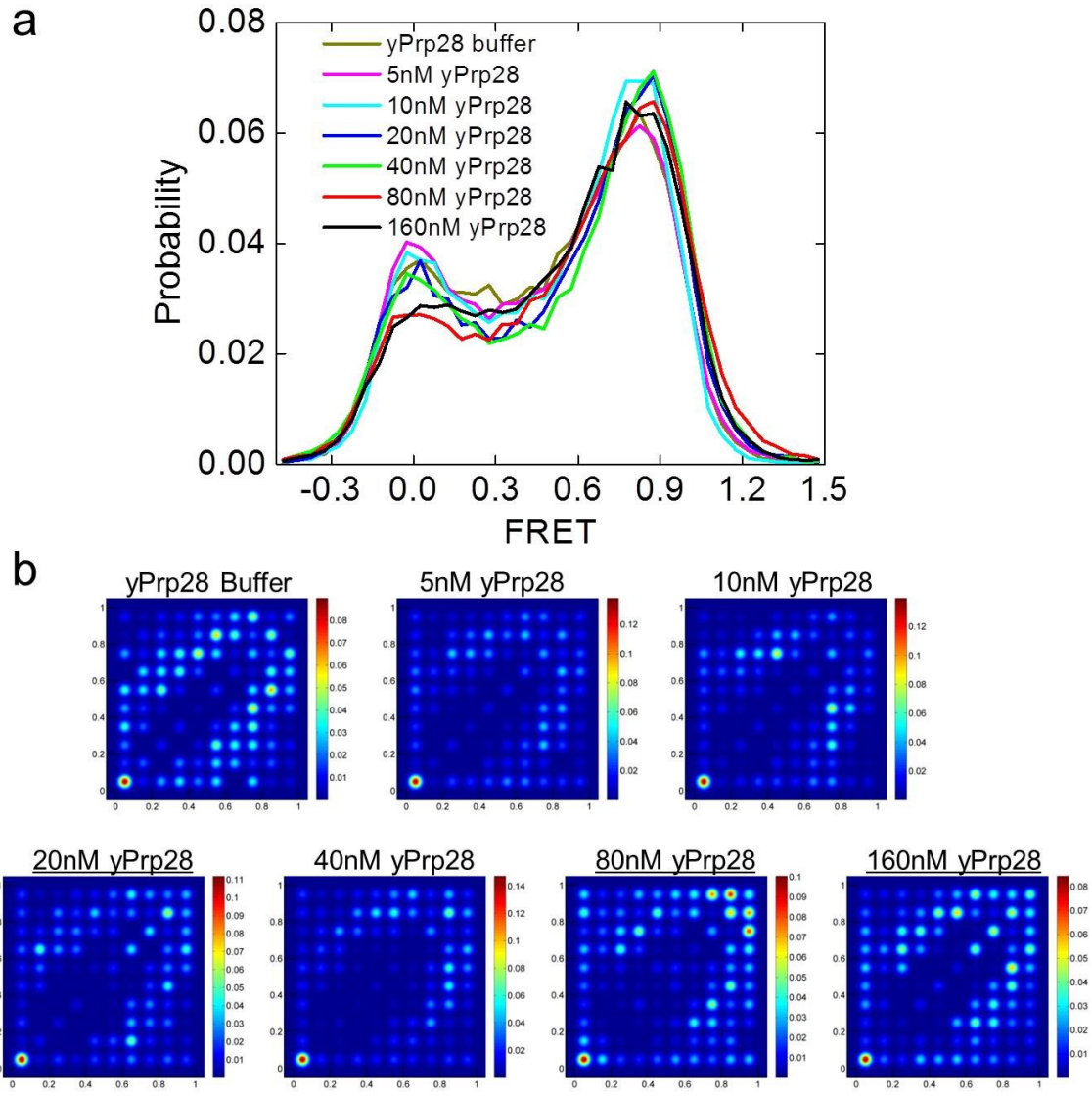



**Figure 6.9 Titration of ΔPrp28 extract with increasing rPrp28 concentrations does not recapitulate previous analysis**

TODP analysis (**Figure B.3b**). All conditions showed a large number of dynamics between the low and high FRET states with no obvious change upon addition of Prp28. These data contradict our initial findings that the depletion of Prp28 results in formation of a 50:50 mixture of low and high FRET conformation so we decided to go back and investigate the initial smFRET analysis.

B.3.4 The BrC mutated substrate also results in formation of a high FRET conformation

Feeling that there may have been a mistake during the initial smFRET investigation of the BP and WT pre-mRNA substrates, we went back to the previously collected smFRET data and re-analyzed both the ΔPrp28-WCE (WT) and ΔPrp28-WCE (BP) datasets. Interestingly, re-selection of suitable Ubc4 molecules containing both Cy3 and Cy5 fluorophores undergoing anti-correlated interconversions and producing a new FRET probability histogram no longer displayed even sized populations of high and low FRET molecules but rather a dominant high FRET peak with both the WT and BP mutant substrate (**Figure B.4a,b**). These data are logical as comparison of the selected molecules revealed a large number of high FRET molecules missing from the initial analysis (**Figure B.4c**).

To further confirm the BP mutant and WT substrates are both characterized by primarily a high FRET conformation when incubated with ΔPrp28-WCE, we repeated these experiments with the same Prp28-depleted extract and pre-mRNA substrates. Once again, as expected, both the mutant and WT pre-mRNA substrate were observed to adopt a primarily high FRET conformation with possibly a slight increase in a low FRET population in the WT substrate over the BP mutant (**Figure B.4a,b**). Given all these data, we concluded that the initial smFRET analysis of the mutant and WT substrates was incorrectly performed and that there is no detectable change in pre-mRNA conformation upon addition of rPrp28 to extract depleted of Prp28.

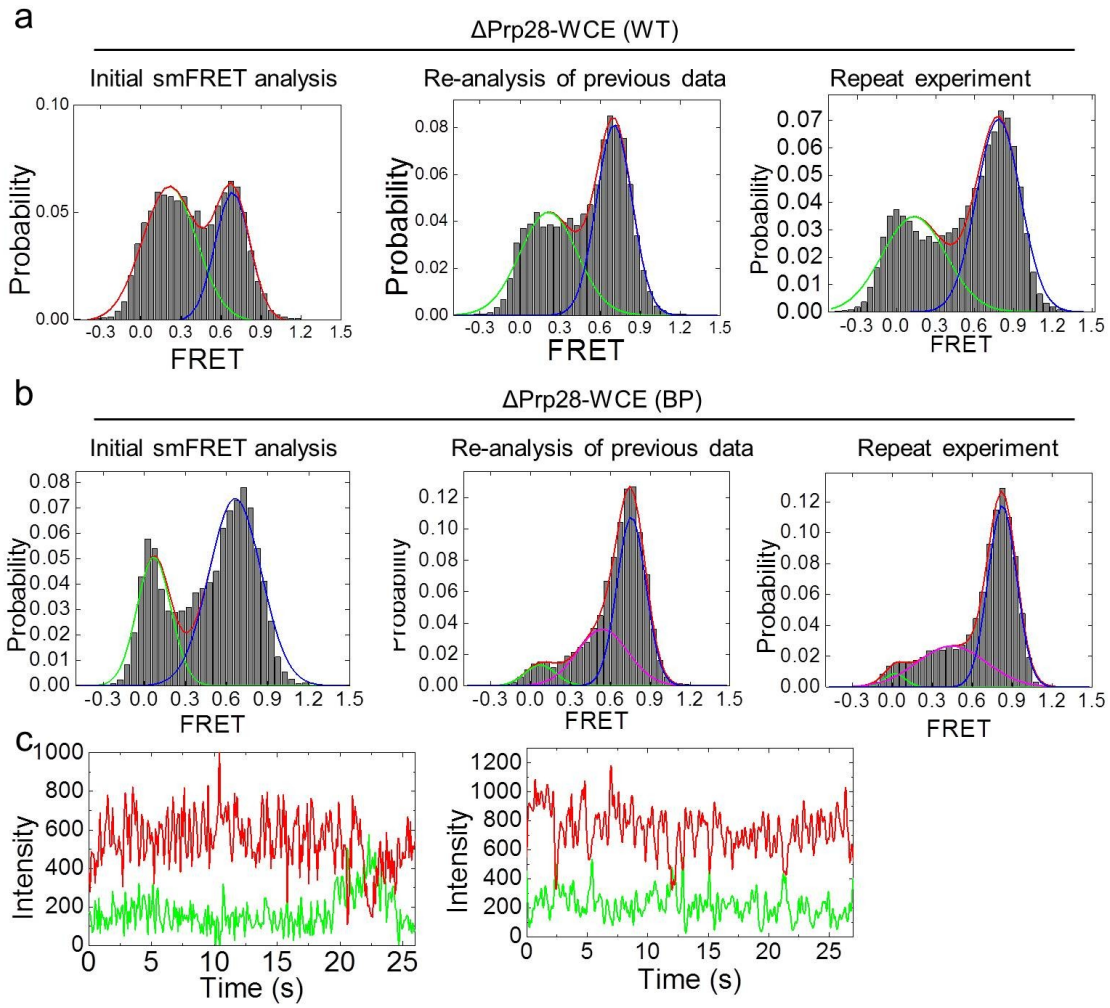



**Figure 6.10 Reanalysis and repeat of the initial experiments reveals mistakes in the analysis**

(a-b) Re-analysis of the initial smFRET results (left) no longer showed an equal population of high and low FRET states but rather a dominant high FRET peak (middle). A repeat experiment further displayed a dominant high FRET population describing both the WT and BP mutant Ubc4 pre-mRNA in the presence of the indicated extract (right). (c) Example single-molecule trajectories showing acceptor (red, Cy5) and donor (green, Cy3) signals that were overlooked in the initial analysis and that might have contributed to the smaller than expected high FRET peak.

## B.4 Discussion

Here we have thoroughly investigated a potential Prp28-dependent change in pre-mRNA conformation early on in spliceosomal commitment complex formation. During the early stages of our analysis, there appeared to be a shift in pre-mRNA conformation towards high FRET upon addition of Prp28 to extracts depleted of Prp28 and ATP. Unfortunately, this result was not reproducible and was actually found to be due to a mistake during molecule selection during the original analysis. Repeat experiments further confirmed that both the BP mutant and WT substrates adopt dominant high FRET conformation in the absence and presence of Prp28. These data support a model in which Prp28 plays a role in BBP/Mud2 stabilization, but binding of the protein dimer to the substrate does not induce a change in pre-mRNA conformation or structure. Interestingly, these data agree with recent computational secondary structure prediction and smFRET data that show Ubc4 adopting a stable 5' stem even in the absence of spliceosomal components[53]. Additionally, the Ubc4 substrate was recently found to adopt a high FRET conformation when extracts were depleted of ATP and a very low, zero FRET state when stalled at the A complex[62]. Considering native gel analysis revealed an equal population of CC1 and CC2, the prevalence of a high FRET conformation would support this being the primary structure for both CC1 and CC2. Furthermore, single-molecule FRET investigation in this thesis supported a high FRET pre-mRNA structure in the absence of yeast extract (Chapter IV). Taken together, these data support formation of a high FRET Ubc4 structure up until A complex formation when binding of the U2 snRNP induces a large change in pre-mRNA conformation in which the BS and 5'SS become separated from one another.

## REFERENCES


1. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).
2. Cech, T.R. & Steitz, J.A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77-94 (2014).
3. Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L. & Steitz, J.A. Are snRNPs involved in splicing? *Nature* **283**, 220-224 (1980).
4. Mello, C.C. & Conte, D., Jr. Revealing the world of RNA interference. *Nature* **431**, 338-342 (2004).
5. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
6. National-Human-Genome-Research-Institute http://www.genome.gov/11006943.
7. Ponting, C.P. & Belgard, T.G. Transcribed dark matter: meaning or myth? *Hum Mol Genet* **19**, R162-168 (2010).
8. Elgar, G. & Vavouri, T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet* **24**, 344-352 (2008).
9. Understanding Our Genetic Inheritance The US Human Genome Project: The First Five Years http://web.ornl.gov/sci/techresources/Human_Genome/project/5yrplan/firstfiveyears.pdf.
10. Saxena, A. & Carninci, P. Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. *Bioessays* **33**, 830-839 (2011).
11. Dinger, M.E., Mercer, T.R. & Mattick, J.S. RNAs as extracellular signaling molecules. *J Mol Endocrinol* **40**, 151-159 (2008).
12. Yoon, J.H. et al. Scaffold function of long non-coding RNA HOTAIR in protein ubiquitination. *Nat Commun* **4**, 2939 (2013).
13. Winkler, W.C., Nahvi, A., Roth, A., Collins, J.A. & Breaker, R.R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**, 281-286 (2004).
14. Jensen, S. & Thomsen, A.R. Sensing of RNA viruses: a review of innate immune receptors involved in recognizing RNA virus invasion. *J Virol* **86**, 2900-2910 (2012).
15. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849-857 (1983).
16. Kruger, K. et al. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147-157 (1982).
17. Marek, M.S., Johnson-Buck, A. & Walter, N.G. The shape-shifting quasispecies of RNA: one sequence, many functional folds. *Phys Chem Chem Phys* **13**, 11524-11537 (2011).
18. Lai, M.M. The molecular biology of hepatitis delta virus. *Annu Rev Biochem* **64**, 259-286 (1995).
19. Hilliker, A.K., Mefford, M.A. & Staley, J.P. U2 toggles iteratively between the stem IIa and stem IIc conformations to promote pre-mRNA splicing. *Genes Dev* **21**, 821-834 (2007).
20. Fica, S.M. et al. RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**, 229-234 (2013).

21. Chow, L.T., Gelinas, R.E., Broker, T.R. & Roberts, R.J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1-8 (1977).
22. Berget, S.M., Moore, C. & Sharp, P.A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**, 3171-3175 (1977).
23. Parenteau, J. et al. Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Mol Biol Cell* **19**, 1932-1941 (2008).
24. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
25. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413-1415 (2008).
26. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476 (2008).
27. Shukla, S. & Oberdoerffer, S. Co-transcriptional regulation of alternative pre-mRNA splicing. *Biochim Biophys Acta* **1819**, 673-683 (2012).
28. Brugiolo, M., Herzel, L. & Neugebauer, K.M. Counting on co-transcriptional splicing. *F1000Prime Rep* **5**, 9 (2013).
29. Hesselberth, J.R. Lives that introns lead after splicing. *Wiley Interdiscip Rev RNA* **4**, 677-691 (2013).
30. Kondo, Y., Oubridge, C., van Roon, A.M. & Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife* **4** (2015).
31. Wang, Q., Zhang, L., Lynn, B. & Rymond, B.C. A BBP-Mud2p heterodimer mediates branchpoint recognition and influences splicing substrate abundance in budding yeast. *Nucleic Acids Res* **36**, 2787-2798 (2008).
32. Kistler, A.L. & Guthrie, C. Deletion of MUD2, the yeast homolog of U2AF65, can bypass the requirement for sub2, an essential spliceosomal ATPase. *Genes Dev* **15**, 42-49 (2001).
33. Kosowski, T.R., Keys, H.R., Quan, T.K. & Ruby, S.W. DExD/H-box Prp5 protein is in the spliceosome during most of the splicing cycle. *RNA* **15**, 1345-1362 (2009).
34. Perriman, R. & Ares, M., Jr. Invariant U2 snRNA nucleotides form a stem loop to recognize the intron early in splicing. *Mol Cell* **38**, 416-427 (2010).
35. Xu, Y.Z. & Query, C.C. Competition between the ATPase Prp5 and branch region-U2 snRNA pairing modulates the fidelity of spliceosome assembly. *Mol Cell* **28**, 838-849 (2007).
36. Will, C.L. & Luhrmann, R. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* **3** (2011).
37. Vijayraghavan, U., Company, M. & Abelson, J. Isolation and characterization of pre-mRNA splicing mutants of Saccharomyces cerevisiae. *Genes Dev* **3**, 1206-1216 (1989).
38. Warkocki, Z. et al. Reconstitution of both steps of Saccharomyces cerevisiae splicing with purified spliceosomal components. *Nat. Struct. Mol. Biol.* **16**, 1237-1243 (2009).
39. Lardelli, R.M., Thompson, J.X., Yates, J.R., 3rd & Stevens, S.W. Release of SF3 from the intron branchpoint activates the first step of pre-mRNA splicing. *RNA* **16**, 516-528 (2010).
40. Ohrt, T. et al. Prp2-mediated protein rearrangements at the catalytic core of the spliceosome as revealed by dcFCCS. *RNA* **18**, 1244-1256 (2012).

41. Tseng, C.K., Liu, H.L. & Cheng, S.C. DEAH-box ATPase Prp16 has dual roles in remodeling of the spliceosome in catalytic steps. *RNA* **17**, 145-154 (2011).
42. Egecioglu, D.E. & Chanfreau, G. Proofreading and spellchecking: a two-tier strategy for pre-mRNA splicing quality control. *RNA* **17**, 383-389 (2011).
43. Mayas, R.M., Maita, H. & Staley, J.P. Exon ligation is proofread by the DExD/H-box ATPase Prp22p. *Nat. Struct. Mol. Biol.* **13**, 482-490 (2006).
44. Förster, T. Zwischenmolekulare Energiewanderung Und Fluoreszenz. *Ann Phys-Berlin* **2**, 55-75 (1948).
45. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat Methods* **5**, 507-516 (2008).
46. Stryer, L. Fluorescence energy transfer as a spectroscopic ruler. *Annu Rev Biochem* **47**, 819-846 (1978).
47. Walter, N.G. Probing RNA structural dynamics and function by fluorescence resonance energy transfer (FRET). *Curr Protoc Nucleic Acid Chem* **Chapter 11**, Unit 11 10 (2003).
48. Walter, N.G. Structural dynamics of catalytic RNA highlighted by fluorescence resonance energy transfer. *Methods* **25**, 19-30 (2001).
49. Walter, N.G., Huang, C.Y., Manzo, A.J. & Sobhy, M.A. Do-it-yourself guide: how to use the modern single-molecule toolkit. *Nat Methods* **5**, 475-489 (2008).
50. Aitken, C.E., Petrov, A. & Puglisi, J.D. Single ribosome dynamics and the mechanism of translation. *Annu Rev Biophys* **39**, 491-513 (2010).
51. Chen, J., Tsai, A., O'Leary, S.E., Petrov, A. & Puglisi, J.D. Unraveling the dynamics of ribosome translocation. *Curr Opin Struct Biol* **22**, 804-814 (2012).
52. Kim, H. et al. Protein-guided RNA dynamics during early ribosome assembly. *Nature* **506**, 334-338 (2014).
53. Abelson, J. et al. Conformational dynamics of single pre-mRNA molecules during in vitro splicing. *Nat Struct Mol Biol* **17**, 504-512 (2010).
54. Crawford, D.J., Hoskins, A.A., Friedman, L.J., Gelles, J. & Moore, M.J. Single-molecule colocalization FRET evidence that spliceosome activation precedes stable approach of 5' splice site and branch site. *Proc Natl Acad Sci U S A* **110**, 6783-6788 (2013).
55. Krishnan, R. et al. Biased Brownian ratcheting leads to pre-mRNA remodeling and capture prior to first-step splicing. *Nat Struct Mol Biol* **20**, 1450-1457 (2013).
56. Hwang, H. et al. Telomeric overhang length determines structural dynamics and accessibility to telomerase and ALT-associated proteins. *Structure* **22**, 842-853 (2014).
57. Parks, J.W. & Stone, M.D. Coordinated DNA dynamics during the human telomerase catalytic cycle. *Nat Commun* **5**, 4146 (2014).
58. Fabrizio, P. et al. The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Mol Cell* **36**, 593-608 (2009).
59. Wahl, M.C., Will, C.L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701-718 (2009).
60. Jain, A. et al. Probing cellular protein complexes using single-molecule pull-down. *Nature* **473**, 484-488 (2011).
61. Jain, A., Liu, R., Xiang, Y.K. & Ha, T. Single-molecule pull-down for studying protein interactions. *Nat Protoc* **7**, 445-452 (2012).
62. Blanco, M.R. et al. Single Molecule Cluster Analysis Identifies Signature Dynamic Conformations along the Splicing Pathway. *Nat Methods* (under revision).

63. Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A. & Weeks, K.M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods* **11**, 959-965 (2014).
64. Brody, E. & Abelson, J. The "spliceosome": yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. *Science* **228**, 963-967 (1985).
65. Staley, J.P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315-326 (1998).
66. Chen, J.H. & Lin, R.J. The yeast PRP2 protein, a putative RNA-dependent ATPase, shares extensive sequence homology with two other pre-mRNA splicing factors. *Nucleic Acids Res.* **18**, 6447 (1990).
67. Berglund, J.A., Rosbash, M. & Schultz, S.C. Crystal structure of a model branchpoint-U2 snRNA duplex containing bulged adenosines. *RNA* **7**, 682-691 (2001).
68. Kim, S.H. & Lin, R.J. Spliceosome activation by PRP2 ATPase prior to the first transesterification reaction of pre-mRNA splicing. *Mol. Cell Biol.* **16**, 6810-6819 (1996).
69. Chiu, Y.F. et al. Cwc25 Is a Novel Splicing Factor Required after Prp2 and Yju2 To Facilitate the First Catalytic Reaction. *Mol Cell Biol* **29**, 5671-5678 (2009).
70. Jain, A. et al. Probing cellular protein complexes using single-molecule pull-down. *Nature* **473**, 484-U322 (2011).
71. Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. *Nature* **425**, 737-741 (2003).
72. 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* **6**, 1106-1119 (2000).
73. Blanco, M. & Walter, N.G. Analysis of complex single-molecule FRET time trajectories. *Methods in enzymology* **472**, 153-178 (2010).
74. Pereira, M.J.B. et al. Single VS ribozyme molecules reveal dynamic and hierarchical folding toward catalysis. *J Mol Biol* **382**, 496-509 (2008).
75. Ditzler, M.A., Rueda, D., Mo, J.J., Hakansson, K. & Walter, N.G. A rugged free energy landscape separates multiple functional RNA folds throughout denaturation. *Nucleic Acids Res.* **36**, 7088-7099 (2008).
76. Sabanayagam, C.R., Eid, J.S. & Meller, A. Using fluorescence resonance energy transfer to measure distances along individual DNA molecules: Corrections due to nonideal transfer. *J Chem Phys* **122** (2005).
77. Cosa, G. et al. Secondary structure and secondary structure dynamics of DNA hairpins complexed with HIV-1 NC protein. *Biophys J* **87**, 2759-2767 (2004).
78. Hwang, H., Kim, H. & Myong, S. Protein induced fluorescence enhancement as a single molecule assay with short distance sensitivity. *Proc Natl Acad Sci U S A* **108**, 7414-7418 (2011).
79. Uphoff, S. et al. Monitoring multiple distances within a single molecule using switchable FRET. *Nature Methods* **7**, 831-U890 (2010).
80. Ishioka, T. Extended K-means with an Efficient Estimation of the Number of Clusters. *Lecture Notes in Computer Science* **1983**, 17-22 (2000).
81. Rueda, D. et al. Single-molecule enzymology of RNA: essential functional groups impact catalysis from a distance. *Proc. Natl. Acad. Sci. USA* **101**, 10066-10071 (2004).
82. Black, D.L. Finding splice sites within a wilderness of RNA. *RNA* **1**, 763-771 (1995).

83. Silverman, E.J. et al. Interaction between a G-patch protein and a spliceosomal DEXD/H-box ATPase that is critical for splicing. *Mol. Cell. Biol.* **24**, 10101-10110 (2004).
84. Roy, J., Kim, K., Maddock, J.R., Anthony, J.G. & Woolford, J.L. The Final Stages of Spliceosome Maturation Require Spp2p That Can Interact with the Deah Box Protein Prp2p and Promote Step-1 of Splicing. *RNA* **1**, 375-390 (1995).
85. Schwer, B. & Guthrie, C. A conformational rearrangement in the spliceosome is dependent on PRP16 and ATP hydrolysis. *EMBO J.* **11**, 5033-5039 (1992).
86. Hotz, H.R. & Schwer, B. Mutational analysis of the yeast DEAH-box splicing factor Prp16. *Genetics* **149**, 807-815 (1998).
87. Teigelkamp, S., McGarvey, M., Plumpton, M. & Beggs, J.D. The splicing factor PRP2, a putative RNA helicase, interacts directly with pre-mRNA. *EMBO J.* **13**, 888-897 (1994).
88. Liu, H.L. & Cheng, S.C. The Interaction of Prp2 with a Defined Region of the Intron Is Required for the First Splicing Reaction. *Mol. Cell. Biol.* **32**, 5056-5066 (2012).
89. Company, M., Arenas, J. & Abelson, J. Requirement of the RNA helicase-like protein PRP22 for release of messenger RNA from spliceosomes. *Nature* **349**, 487-493 (1991).
90. Newnham, C.M. & Query, C.C. The ATP requirement for U2 snRNP addition is linked to the pre-mRNA region 5 ' to the branch site. *RNA* **7**, 1298-1309 (2001).
91. Perriman, R., Barta, I., Voeltz, G.K., Abelson, J. & Ares, M. ATP requirement for Prp5p function is determined by Cus2p and the structure of U2 small nuclear RNA. *Proc. Natl. Acad. Sci. USA* **100**, 13857-13862 (2003).
92. Schwer, B. & Gross, C.H. Prp22, a DExH-box RNA helicase, plays two distinct roles in yeast pre-mRNA splicing. *EMBO J.* **17**, 2086-2094 (1998).
93. Schwer, B. A conformational rearrangement in the spliceosome sets the stage for prp22-dependent mRNA release. *Mol. Cell* **30**, 743-754 (2008).
94. Bartels, C., Klatt, C., Luhrmann, R. & Fabrizio, P. The ribosomal translocase homologue Snu114p is involved in unwinding U4/U6 RNA during activation of the spliceosome. *Embo Rep* **3**, 875-880 (2002).
95. Raghunathan, P.L. & Guthrie, C. RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2. *Curr. Biol.* **8**, 847-855 (1998).
96. Kim, S.H. The purified yeast pre-mRNA splicing factor PRP2 is an RNA-dependent NTPase. *EMBO J.* **11**, 2319–2326 (1992).
97. Ohi, M.D. et al. Proteomics analysis reveals stable multiprotein complexes in both fission and budding yeasts containing Myb-related Cdc5p/Cef1p, novel pre-mRNA splicing factors, and snRNAs. *Mol. Cell. Biol.* **22**, 2011-2024 (2002).
98. Chen, H.C., Tseng, C.K., Tsai, R.T., Chung, C.S. & Cheng, S.C. Link of NTR-mediated spliceosome disassembly with DEAH-box ATPases Prp2, Prp16, and Prp22. *Mol Cell Biol* **33**, 514-525 (2013).
99. Cordova, N.J., Ermentrout, B. & Oster, G.F. Dynamics of single-motor molecules: the thermal ratchet model. *Proc. Natl. Acad. Sci. USA* **89**, 339-343 (1992).
100. Astumian, R.D. Thermodynamics and kinetics of a Brownian motor. *Science* **276**, 917-922 (1997).
101. Spirin, A.S. How Does a Scanning Ribosomal Particle Move along the 5 '-Untranslated Region of Eukaryotic mRNA? Brownian Ratchet Model. *Biochemistry-Us* **48**, 10688-10692 (2009).
102. Frank, J. & Gonzalez, R.L., Jr. Structure and dynamics of a processive Brownian motor: the translating ribosome. *Annu. Rev. Biochem.* **79**, 381-412 (2010).

103. Rodnina, M.V. & Wintermeyer, W. The ribosome as a molecular machine: the mechanism of tRNA-mRNA movement in translocation. *Biochem. Soc. Trans.* **39**, 658-662 (2011).
104. Zhou, J., Lancaster, L., Donohue, J.P. & Noller, H.F. Crystal structures of EF-G-ribosome complexes trapped in intermediate states of translocation. *Science* **340**, 1236086 (2013).
105. Blanchard, S.C., Gonzalez, R.L., Kim, H.D., Chu, S. & Puglisi, J.D. tRNA selection and kinetic proofreading in translation. *Nat Struct Mol Biol* **11**, 1008-1014 (2004).
106. Tanner, N.K. & Linder, P. DExD/H box RNA helicases: from generic motors to specific dissociation functions. *Mol. Cell* **8**, 251-262 (2001).
107. Jankowsky, E. RNA helicases at work: binding and rearranging. *Trends Biochem. Sci.* **36**, 19-29 (2011).
108. Robertson, K.L., Yu, L., Armitage, B.A., Lopez, A.J. & Peteanu, L.A. Fluorescent PNA probes as hybridization labels for biological RNA. *Biochemistry-Us* **45**, 6066-6074 (2006).
109. Crawford, D.J., Hoskins, A.A., Friedman, L.J., Gelles, J. & Moore, M.J. Visualizing the splicing of single pre-mRNA molecules in whole cell extract. *RNA* **14**, 170-179 (2008).
110. Hoskins, A.A. et al. Ordered and dynamic assembly of single spliceosomes. *Science* **331**, 1289-1295 (2011).
111. Pitchiaya, S., Heinicke, L.A., Custer, T.C. & Walter, N.G. Single molecule fluorescence approaches shed light on intracellular RNAs. *Chemical reviews* **114**, 3224-3265 (2014).
112. Mustoe, A.M., Brooks, C.L. & Al-Hashimi, H.M. Hierarchy of RNA functional dynamics. *Annu. Rev. Biochem.* **83**, 441-466 (2014).
113. Al-Hashimi, H.M. & Walter, N.G. RNA dynamics: it is about time. *Curr. Opin. Struct. Biol.* **18**, 321-329 (2008).
114. Cruz, J.A. & Westhof, E. The dynamic landscapes of RNA architecture. *Cell* **136**, 604-609 (2009).
115. Walter, N.G. & Bustamante, C. Introduction to single molecule imaging and mechanics: seeing and touching molecules one at a time. *Chemical reviews* **114**, 3069-3071 (2014).
116. Semlow, D.R. & Staley, J.P. Staying on message: ensuring fidelity in pre-mRNA splicing. *Trends Biochem Sci* **37**, 263-273 (2012).
117. Abelson, J., Hadjivassiliou, H. & Guthrie, C. Preparation of fluorescent pre-mRNA substrates for an smFRET study of pre-mRNA splicing in yeast. *Methods in enzymology* **472**, 31-40 (2010).
118. Gopich, I.V. & Szabo, A. Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET. *Proc Natl Acad Sci U S A* **109**, 7747-7752 (2012).
119. Greenfeld, M., Pavlichin, D.S., Mabuchi, H. & Herschlag, D. Single Molecule Analysis Research Tool (SMART): an integrated approach for analyzing single molecule data. *PLoS One* **7**, e30024 (2012).
120. Keller, B.G., Kobitski, A., Jaschke, A., Nienhaus, G.U. & Noe, F. Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models. *J Am Chem Soc* **136**, 4534-4543 (2014).
121. Stevens, S.W. & Abelson, J. Yeast pre-mRNA splicing: methods, mechanisms, and machinery. *Methods in enzymology* **351**, 200-220 (2002).

122. Widom, J.R., Dhakal, S., Heinicke, L.A. & Walter, N.G. Single molecule tools for enzymology, structural biology, systems biology and nanotechnology: an update. *Arch. Toxicol.* **in press** (2014).
123. Bronson, J.E., Fei, J., Hofman, J.M., Gonzalez, R.L., Jr. & Wiggins, C.H. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys J* **97**, 3196-3205 (2009).
124. Mall, R., Langone, R. & Suykens, J.A. Multilevel hierarchical kernel spectral clustering for real-life large scale complex networks. *PLoS One* **9**, e99966 (2014).
125. Bruno, A.E. et al. Comparing chemistry to outcome: the development of a chemical distance metric, coupled with clustering and hierarchal visualization applied to macromolecular crystallography. *PLoS One* **9**, e100782 (2014).
126. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. Royal Stat. Soc. Ser. B* **63**, 411-423 (2001).
127. Koodathingal, P., Novak, T., Piccirilli, J.A. & Staley, J.P. The DEAH box ATPases Prp16 and Prp43 cooperate to proofread 5' splice site cleavage during pre-mRNA splicing. *Mol Cell* **39**, 385-395 (2010).
128. Schneider, S., Hotz, H.R. & Schwer, B. Characterization of dominant-negative mutants of the DEAH-box splicing factors Prp22 and Prp16. *J Biol Chem* **277**, 15452-15458 (2002).
129. Ohrt, T. et al. Molecular dissection of step 2 catalysis of yeast pre-mRNA splicing investigated in a purified system. *RNA* **19**, 902-915 (2013).
130. Umen, J.G. & Guthrie, C. Prp16p, Slu7p, and Prp8p interact with the 3' splice site in two distinct stages during the second catalytic step of pre-mRNA splicing. *RNA* **1**, 584-597 (1995).
131. Woese, C.R. & Fox, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74**, 5088-5090 (1977).
132. McGinnis, J.L., Duncan, C.D. & Weeks, K.M. High-throughput SHAPE and hydroxyl radical analysis of RNA structure and ribonucleoprotein assembly. *Methods in enzymology* **468**, 67-89 (2009).
133. Das, R., Laederach, A., Pearlman, S.M., Herschlag, D. & Altman, R.B. SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA* **11**, 344-354 (2005).
134. Suddala, K.C. et al. Single transcriptional and translational preQ1 riboswitches adopt similar pre-folded ensembles that follow distinct folding pathways into the same ligand-bound structure. *Nucleic Acids Res* **41**, 10462-10475 (2013).
135. Suddala, K.C. & Walter, N.G. Riboswitch structure and dynamics by smFRET microscopy. *Methods in enzymology* **549**, 343-373 (2014).
136. Qin, P.Z. & Pyle, A.M. Site-specific labeling of RNA with fluorophores and other structural probes. *Methods* **18**, 60-70 (1999).
137. Jeong, S., Sefcikova, J., Tinsley, R.A., Rueda, D. & Walter, N.G. Trans-acting hepatitis delta virus ribozyme: catalytic core and global structure are dependent on the 5' substrate sequence. *Biochemistry-Us* **42**, 7727-7740 (2003).
138. Harris, D.A., Tinsley, R.A. & Walter, N.G. Terbium-mediated footprinting probes a catalytic conformational switch in the antigenomic hepatitis delta virus ribozyme. *J Mol Biol* **341**, 389-403 (2004).

139. Walter, N.G., Yang, N. & Burke, J.M. Probing non-selective cation binding in the hairpin ribozyme with Tb(III). *J Mol Biol* **298**, 539-555 (2000).
140. Sefcikova, J., Krasovska, M.V., Spackova, N., Sponer, J. & Walter, N.G. Impact of an extruded nucleotide on cleavage activity and dynamic catalytic core conformation of the hepatitis delta virus ribozyme. *Biopolymers* **85**, 392-406 (2007).
141. Rabiner, L.R. A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition. *Proceedings of the IEEE* **77**, 257-286 (1989).
142. Kucharik, M., Hofacker, I.L., Stadler, P.F. & Qin, J. Basin Hopping Graph: a computational framework to characterize RNA folding landscapes. *Bioinformatics* **30**, 2009-2017 (2014).
143. Tinoco, I., Jr. et al. Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol* **246**, 40-41 (1973).
144. Hofacker, I.L., Schuster, P. & Stadler, P.F. Combinatorics of RNA secondary structures. *Discrete Applied Mathematics* **88**, 207-237 (1998).
145. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003).
146. Crippen, G.M. & Havel, T.F. Distance geometry and molecular conformation. *Research Studies Press, Taunton, England* (1988).
147. Havel, T.F. An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog Biophys Mol Biol* **56**, 43-78 (1991).
148. Ren, P. & Ponder, J.W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation *J. Phys. Chem. B* **107**, 5933-5947 (2003).
149. Sawa, H. & Abelson, J. Evidence for a base-pairing interaction between U6 small nuclear RNA and 5' splice site during the splicing reaction in yeast. *Proc Natl Acad Sci U S A* **89**, 11269-11273 (1992).
150. Kandels-Lewis, S. & Seraphin, B. Involvement of U6 snRNA in 5' splice site selection. *Science* **262**, 2035-2039 (1993).
151. Yu, A.T., Ge, J. & Yu, Y.T. Pseudouridines in spliceosomal snRNAs. *Protein Cell* **2**, 712-725 (2011).
152. Berglund, J.A., Chua, K., Abovich, N., Reed, R. & Rosbash, M. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell* **89**, 781-787 (1997).
153. Wang, G.S. & Cooper, T.A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**, 749-761 (2007).
154. Goguel, V. & Rosbash, M. Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. *Cell* **72**, 893-901 (1993).
155. Toor, N., Keating, K.S. & Pyle, A.M. Structural insights into RNA splicing. *Curr Opin Struct Biol* **19**, 260-266 (2009).
156. Rogic, S. et al. Correlation between the secondary structure of pre-mRNA introns and the efficiency of splicing in Saccharomyces cerevisiae. *BMC Genomics* **9**, 355 (2008).
157. Meyer, M., Plass, M., Perez-Valle, J., Eyras, E. & Vilardell, J. Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol Cell* **43**, 1033-1039 (2011).
158. Shen, M. et al. Pyrvinium pamoate changes alternative splicing of the serotonin receptor 2C by influencing its RNA structure. *Nucleic Acids Res* **41**, 3819-3832 (2013).

159. Paulsen, M.T. et al. Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods* **67**, 45-54 (2014).
160. Paulsen, M.T. et al. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc Natl Acad Sci U S A* **110**, 2240-2245 (2013).
161. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578 (2012).
162. Pleiss, J.A., Whitworth, G.B., Bergkessel, M. & Guthrie, C. Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biol* **5**, e90 (2007).
163. Miura, F. et al. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A* **103**, 17846-17851 (2006).
164. Volanakis, A. et al. Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast. *Genes Dev* **27**, 2025-2038 (2013).
165. Cherry, J.M. et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**, D700-705 (2012).
166. Ares, M. http://intron.ucsc.edu/yeast4.3/. (2011).
167. Zhang, D., Abovich, N. & Rosbash, M. A biochemical function for the Sm complex. *Mol Cell* **7**, 319-329 (2001).
168. Juneau, K., Palm, C., Miranda, M. & Davis, R.W. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *Proc Natl Acad Sci U S A* **104**, 1522-1527 (2007).
169. Munding, E.M. et al. Integration of a splicing regulatory network within the meiotic gene expression program of Saccharomyces cerevisiae. *Genes Dev* **24**, 2693-2704 (2010).
170. Kim Guisbert, K.S. et al. Meiosis-induced alterations in transcript architecture and noncoding RNA expression in S. cerevisiae. *RNA* **18**, 1142-1153 (2012).
171. Brar, G.A. et al. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**, 552-557 (2012).
172. Qi, X. et al. Prevalent and distinct spliceosomal 3'-end processing mechanisms for fungal telomerase RNA. *Nat Commun* **6**, 6105 (2015).
173. Huh, W.K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686-691 (2003).
174. Albulescu, L.O. et al. A quantitative, high-throughput reverse genetic screen reveals novel connections between Pre-mRNA splicing and 5' and 3' end transcript determinants. *PLoS Genet* **8**, e1002530 (2012).
175. Mortimer, S.A. & Weeks, K.M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J Am Chem Soc* **129**, 4144-4145 (2007).
176. Stark, H. & Luhrmann, R. Cryo-electron microscopy of spliceosomal components. *Annu Rev Biophys Biomol Struct* **35**, 435-457 (2006).
177. Golas, M.M. et al. 3D cryo-EM structure of an active step I spliceosome and localization of its catalytic core. *Mol Cell* **40**, 927-938 (2010).
178. Sander, B. et al. Organization of core spliceosomal components U5 snRNA loop I and U4/U6 Di-snRNP within U4/U6.U5 Tri-snRNP as revealed by electron cryomicroscopy. *Mol Cell* **24**, 267-278 (2006).

179. Hacker, I. et al. Localization of Prp8, Brr2, Snu114 and U4/U6 proteins in the yeast tri-snRNP by electron microscopy. *Nat Struct Mol Biol* **15**, 1206-1212 (2008).
180. Karaduman, R. et al. Structure of yeast U6 snRNPs: arrangement of Prp24p and the LSm complex as revealed by electron microscopy. *RNA* **14**, 2528-2537 (2008).
181. Nguyen, T.H. et al. Structural basis of Brr2-Prp8 interactions and implications for U5 snRNP biogenesis and the spliceosome active site. *Structure* **21**, 910-919 (2013).
182. Galej, W.P., Oubridge, C., Newman, A.J. & Nagai, K. Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature* **493**, 638-643 (2013).
183. Wolf, E., Kastner, B. & Luhrmann, R. Antisense-targeted immuno-EM localization of the pre-mRNA path in the spliceosomal C complex. *RNA* **18**, 1347-1357 (2012).
184. Wolf, E. et al. Exon, intron and splice site locations in the spliceosomal B complex. *EMBO J* **28**, 2283-2292 (2009).
185. Matera, A.G. & Wang, Z. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol* **15**, 108-121 (2014).
186. Cartegni, L., Chew, S.L. & Krainer, A.R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3**, 285-298 (2002).
187. Poulos, M.G., Batra, R., Charizanis, K. & Swanson, M.S. Developments in RNA splicing and disease. *Cold Spring Harb Perspect Biol* **3**, a000778 (2011).
188. David, C.J. & Manley, J.L. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev* **24**, 2343-2364 (2010).
189. Xiong, H.Y. et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
190. Semlow, D.R., Blanco, M., Kahlscheuer, M.L., Walter, N.G. & Staley, J.P. Spliceosomal DExD/H-box ATPases transiently separate candidate splice sites and enable splice site selection. *Cell*, under revision (2015).
191. Buttner, L., Javadi-Zarnaghi, F. & Hobartner, C. Site-specific labeling of RNA at internal ribose hydroxyl groups: terbium-assisted deoxyribozymes at work. *J Am Chem Soc* **136**, 8131-8137 (2014).
192. McPheeters, D.S., Fabrizio, P. & Abelson, J. In vitro reconstitution of functional yeast U2 snRNPs. *Genes Dev* **3**, 2124-2136 (1989).
193. Mozaffari-Jovin, S. et al. The Prp8 RNase H-like domain inhibits Brr2-mediated U4/U6 snRNA unwinding by blocking Brr2 loading onto the U4 snRNA. *Genes Dev* **26**, 2422-2434 (2012).
194. Tanaka, N., Aronova, A. & Schwer, B. Ntr1 activates the Prp43 helicase to trigger release of lariat-intron from the spliceosome. *Genes Dev* **21**, 2312-2325 (2007).
195. Wang, Y., Wagner, J.D. & Guthrie, C. The DEAH-box splicing factor Prp16 unwinds RNA duplexes in vitro. *Curr Biol* **8**, 441-451 (1998).
196. Syed, S., Pandey, M., Patel, S.S. & Ha, T. Single-molecule fluorescence reveals the unwinding stepping mechanism of replicative helicase. *Cell Rep* **6**, 1037-1045 (2014).
197. Prescher, J.A. & Bertozzi, C.R. Chemistry in living systems. *Nat Chem Biol* **1**, 13-21 (2005).
198. Volkmann, G. & Liu, X.Q. Protein C-terminal labeling and biotinylation using synthetic peptide and split-intein. *PLoS One* **4**, e8381 (2009).
199. Shi, X. et al. Quantitative fluorescence labeling of aldehyde-tagged proteins for single-molecule imaging. *Nat Methods* **9**, 499-503 (2012).

200. Chattopadhaya, S., Abu Bakar, F.B. & Yao, S.Q. Expanding the chemical biologist's tool kit: chemical labelling strategies and its applications. *Curr Med Chem* **16**, 4527-4543 (2009).
201. Vila-Perello, M. & Muir, T.W. Biological applications of protein splicing. *Cell* **143**, 191-200 (2010).
202. Volkmann, G. & Iwai, H. Protein trans-splicing and its use in structural biology: opportunities and limitations. *Mol Biosyst* **6**, 2110-2121 (2010).
203. Gogarten, J.P., Senejani, A.G., Zhaxybayeva, O., Olendzenski, L. & Hilario, E. Inteins: structure, function, and evolution. *Annu Rev Microbiol* **56**, 263-287 (2002).
204. Appleby, J.H., Zhou, K., Volkmann, G. & Liu, X.Q. Novel split intein for trans-splicing synthetic peptide onto C terminus of protein. *J Biol Chem* **284**, 6194-6199 (2009).
205. Frese, M.A. & Dierks, T. Formylglycine aldehyde Tag--protein engineering through a novel post-translational modification. *Chembiochem* **10**, 425-427 (2009).
206. Carlson, B.L. et al. Function and structure of a prokaryotic formylglycine-generating enzyme. *J Biol Chem* **283**, 20117-20125 (2008).
207. Dierks, T. et al. Molecular basis for multiple sulfatase deficiency and mechanism for formylglycine generation of the human formylglycine-generating enzyme. *Cell* **121**, 541-552 (2005).
208. Carrico, I.S., Carlson, B.L. & Bertozzi, C.R. Introducing genetically encoded aldehydes into proteins. *Nat Chem Biol* **3**, 321-322 (2007).
209. Wu, P. et al. Site-specific chemical modification of recombinant proteins produced in mammalian cells by using the genetically encoded aldehyde tag. *Proc Natl Acad Sci U S A* **106**, 3000-3005 (2009).
210. Zanetti-Domingues, L.C., Tynan, C.J., Rolfe, D.J., Clarke, D.T. & Martin-Fernandez, M. Hydrophobic fluorescent probes introduce artifacts into single molecule tracking experiments due to non-specific binding. *PLoS One* **8**, e74200 (2013).
211. Lancia, J.K. et al. Sequence context and crosslinking mechanism affect the efficiency of in vivo capture of a protein-protein interaction. *Biopolymers* **101**, 391-397 (2014).
212. Krishnamurthy, M. et al. Caught in the act: covalent cross-linking captures activator-coactivator interactions in vivo. *ACS Chem Biol* **6**, 1321-1326 (2011).
213. Majmudar, C.Y. et al. Impact of nonnatural amino acid mutagenesis on the in vivo function and binding modes of a transcriptional activator. *J Am Chem Soc* **131**, 14240-14242 (2009).
214. Price, A.M., Gornemann, J., Guthrie, C. & Brow, D.A. An unanticipated early function of DEAD-box ATPase Prp28 during commitment to splicing is modulated by U5 snRNP protein Prp8. *RNA* **20**, 46-60 (2014).
215. Seraphin, B. & Rosbash, M. Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. *Cell* **59**, 349-358 (1989).
216. Rutz, B. & Seraphin, B. Transient interaction of BBP/ScSF1 and Mud2 with the splicing machinery affects the kinetics of spliceosome assembly. *RNA* **5**, 819-831 (1999).
217. Abovich, N. & Rosbash, M. Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell* **89**, 403-412 (1997).