CHAPTER NINE

Putting Humpty–Dumpty Together: Clustering the Functional Dynamics of Single Biomolecular Machines Such as the Spliceosome

C.E. Rohlman[*], **M.R. Blanco**[†,1], **N.G. Walter**[†,2]
[*]Albion College, Albion, MI, United States
[†]Single Molecule Analysis Group and Center for RNA Biomedicine, University of Michigan, Ann Arbor, MI, United States
[2]Corresponding author: e-mail address: nwalter@umich.edu

Contents

Abstract

The spliceosome is a biomolecular machine that, in all eukaryotes, accomplishes site-specific splicing of introns from precursor messenger RNAs (pre-mRNAs) with high fidelity. Operating at the nanometer scale, where inertia and friction have lost the dominant role they play in the macroscopic realm, the spliceosome is highly dynamic and assembles its active site around each pre-mRNA anew. To understand the structural dynamics underlying the molecular motors, clocks, and ratchets that achieve functional accuracy in the yeast spliceosome (a long-standing model system), we have developed single-molecule fluorescence resonance energy transfer (smFRET) approaches that report changes in intra- and intermolecular interactions in real time. Building on our work using

[1] Current address: California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA.

Methods in Enzymology, Volume 581
ISSN 0076-6879
http://dx.doi.org/10.1016/bs.mie.2016.08.022

hidden Markov models (HMMs) to extract kinetic and conformational state information from smFRET time trajectories, we recognized that HMM analysis of individual state transitions as independent stochastic events is insufficient for a biomolecular machine as complex as the spliceosome. In this chapter, we elaborate on the recently developed smFRET-based Single-Molecule Cluster Analysis (SiMCAn) that dissects the intricate conformational dynamics of a pre-mRNA through the splicing cycle in a model-free fashion. By leveraging hierarchical clustering techniques developed for Bioinformatics, SiMCAn efficiently analyzes large datasets to first identify common molecular behaviors. Through a second level of clustering based on the abundance of dynamic behaviors exhibited by defined functional intermediates that have been stalled by biochemical or genetic tools, SiMCAn then efficiently assigns pre-mRNA FRET states and transitions to specific splicing complexes, with the potential to find heretofore undescribed conformations. SiMCAn thus arises as a general tool to analyze dynamic cellular machines more broadly.

1. INTRODUCTION

In all eukaryotes, precursor messenger RNA (pre-mRNA) splicing, i.e., the removal of introns and ligation of flanking exons, is a crucial step in the maturation of an mRNA. In higher eukaryotes such as humans, alternative splicing allows for cell- and tissue specific, tightly regulated expression of diverse protein isoforms from a single gene (de Klerk & t Hoen, 2015; Nilsen & Graveley, 2010; Papasaikas & Valcarcel, 2016). A vast majority (~94%) of all human genes contain introns with an average size of 1000–2000 nucleotides (nt), approximately 10 times the size of the protein-coding exons (Barash et al., 2010; Braunschweig, Gueroussov, Plocik, Graveley, & Blencowe, 2013; Calarco, Zhen, & Blencowe, 2011; Chen & Manley, 2009; de Klerk & t Hoen, 2015; Kornblihtt et al., 2013; Lander et al., 2001; Papasaikas & Valcarcel, 2016; Rino & Carmo-Fonseca, 2009). At multimegadalton, perhaps the largest and most elaborate macromolecular machine of the cell (Chen & Moore, 2014; Matera & Wang, 2014; Nilsen, 2003; Papasaikas & Valcarcel, 2016), the spliceosome is responsible for most cellular splicing processes. It must reliably identify intron–exon boundaries, precisely ligate adjacent exons without inadvertent skipping, and appropriately regulate alternative splicing to meet the biological requirements of each cell independent of its environment (Kornblihtt et al., 2013; Matera & Wang, 2014; Naftelberg, Schor, Ast, & Kornblihtt, 2015; Papasaikas & Valcarcel, 2016; Wang & Cooper, 2007). Intron excision with single-nucleotide precision is thus no easy task. Not surprisingly, up to 50% of all mutations leading to human disease are thought to originate from splicing defects (Cartegni, Chew, & Krainer, 2002; Kornblihtt et al., 2013;

Li et al., 2016; Matera & Wang, 2014; Naftelberg et al., 2015; Poulos, Batra, Charizanis, & Swanson, 2011; Scotti & Swanson, 2016; Wang & Cooper, 2007; Zhang & Manley, 2013).

The spliceosome lacks a preformed catalytic core. Rather, for each splicing cycle a spliceosome is—just like the proverbial Humpty–Dumpty—put together anew on an intron-containing pre-mRNA substrate in a stepwise manner that requires both binding and release of small nuclear ribonucleoproteins (snRNPs), each formed by protein cofactors binding to a single small nuclear RNA (snRNA) (Fig. 1; Matera & Wang, 2014; Papasaikas & Valcarcel, 2016; Wahl, Will, & Luhrmann, 2009; Will & Luhrmann, 2011). Notably, splicing entails two successive phosphoryl transfer reactions that are isoenergetic and have been shown to also be reversible, at least under a certain set of ionic conditions (Smith & Konarska, 2008; Tseng & Cheng, 2008). Nonetheless, spliceosome assembly requires at least eight RNA-dependent ATPases of the so-called DExD/H-box subfamily (Jarmoskaite & Russell, 2014; Koodathingal & Staley, 2013; Staley & Guthrie, 1998). These enzymes, which catalyze successive rearrangements of RNA and protein components, are believed to enhance the fidelity of splicing by acting as proofreading clocks that allow the spliceosome to discard mutant substrates at multiple steps in the assembly and catalytic pathway (Burgess & Guthrie, 1993; Couto, Tamm, Parker, & Guthrie, 1987; De, Schmitzova, & Pena, 2016; Horowitz, 2011; Jarmoskaite & Russell, 2014; Koodathingal, Novak, Piccirilli, & Staley, 2010; Koodathingal & Staley, 2013; Mayas, Maita, Semlow, & Staley, 2010; Mayas, Maita, & Staley, 2006; Xu & Query, 2007).

For the budding yeast *Saccharomyces cerevisiae*, a vast array of tools for genetic and biochemical manipulation are available, making this simple, unicellular eukaryote an ideal model organism to study many biological processes. Much has been learned about the dynamic yeast spliceosome since its discovery (Fig. 1; Chen & Moore, 2014; Matera & Wang, 2014; Papasaikas & Valcarcel, 2016; Wahl et al., 2009; Will & Luhrmann, 2011). Assembly is initiated by the ATP-independent formation of a commitment complex in which the U1 snRNP interacts with the 5′ splice site (5′SS), while BBP and Mud2 interact with the branchpoint (BP) sequence. If the BP is mutated (UACUAAC → UACUACC), this complex cannot form and spliceosome assembly is blocked. With a wild-type (WT) precursor, the commitment complex is converted to the prespliceosome by the ATP-dependent binding of U2 to the BP, followed by association of the U4/U6.U5 triple snRNP and Prp19 complexes. Before the first

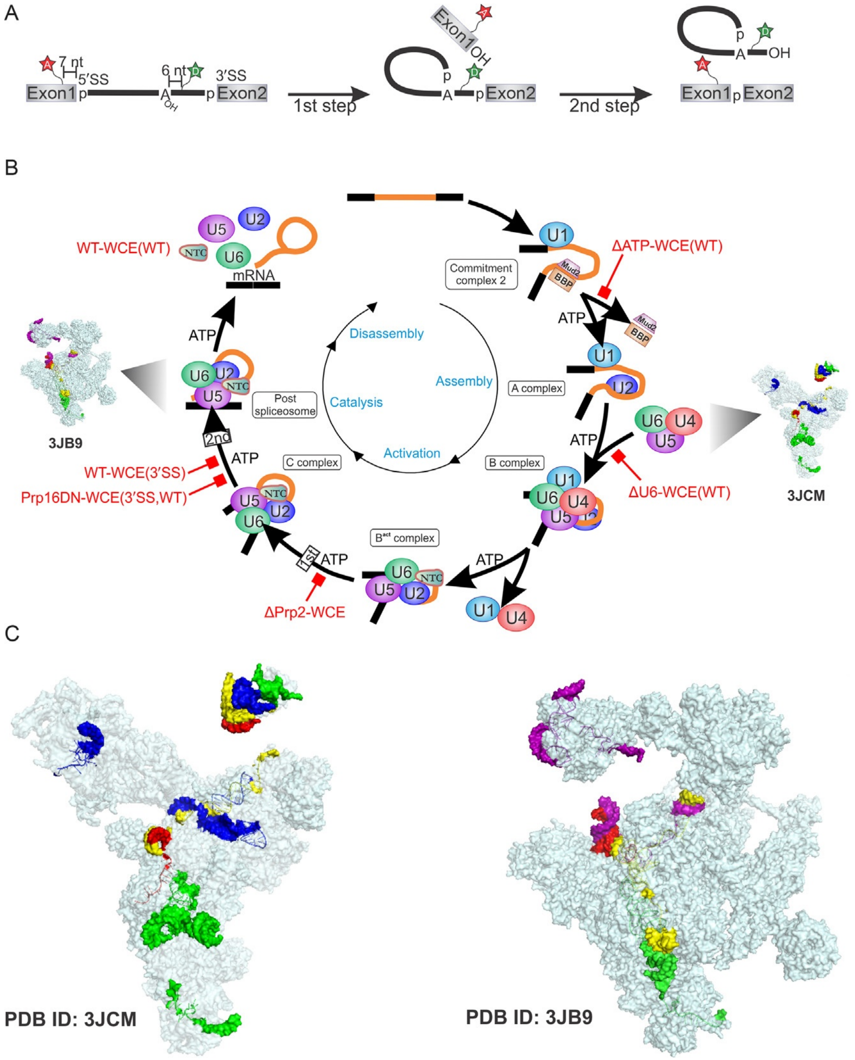

Fig. 1 The pre-mRNA splicing cycle. (A) The Ubc4 pre-mRNA substrate used to monitor pre-mRNA dynamics by smFRET contains Cy5 and Cy3 fluorophores seven nucleotides upstream of the 5′SS and six nucleotides downstream of the BP, respectively. (B) The spliceosome assembly and catalysis pathway is thought to progress in a stepwise manner requiring ATP at several steps of assembly. The biochemical and genetic stalls referred to in this chapter are indicated by *red blocks*. (C) Molecular models of spliceosomal intermediates (with more currently emerging). Solvent accessible surface map of spliceosomal intermediates based on reported cryo-EM structures (PDB IDs 3JCM and 3JB9). RNA components are highlighted in *red* for substrate or product RNA (pre-mRNA or lariat), *blue* for U4 snRNA, *yellow* for U6 snRNA, *green* for U5 snRNA, and *purple* for U2 snRNA. Images were generated in PyMol (DeLano, 2002).

transesterification can occur, U1 and U4 snRNPs must be released from the assembled spliceosome to form the activated B^{act} complex, again in ATP-dependent rearrangements. The DExD/H-box helicase Prp2 then converts this complex to the catalytically active B^{act} complex by destabilizing the U2 snRNP protein complexes (SF3a and SF3b) such that the BP adenosine becomes available for a nucleophilic attack on the distal 5′SS (Fig. 1B). This first chemical step, which results in formation of the looped lariat intermediate and free 5′ exon, can occur even in the presence of a mutation at the 3′ splice site (3′SS mutant UAG → UAC) known to block the second transesterification. After the Prp16- and Prp22-dependent second chemical step of splicing, the spliceosome undergoes ATP-dependent disassembly from the lariat intron and mature mRNA (Fig. 1A and B).

Recent advances in X-ray crystallography and particularly cryogenic electron microscopy (cryo-EM) have started to reveal near-atomic resolution structures of three select spliceosomal complexes from budding (Nguyen et al., 2015; Nguyen, Galej, Fica, et al., 2016; Wan et al., 2016) and fission yeast (Hang, Wan, Yan, & Shi, 2015; Yan et al., 2015), as well as from human (Agafonov et al., 2016; Pomeranz Krummel, Oubridge, Leung, Li, & Nagai, 2009), with the hope that more gaps around the splicing cycle may soon be filled (Fig. 1 and Table 1; Nguyen, Galej, Fica, et al., 2016). Despite three decades of study leading to these insights, however, until recently there was still precious little known about the timing and coordination of the multiple compositional and conformational rearrangements, which together constitute its enzymatic mechanism and endow the spliceosome with high efficiency and fidelity. This gap in knowledge was in large part due to, first, a paucity of biophysical tools suited to

Table 1 Currently Available High-Resolution Cryo-EM Yeast Spliceosome Structures

PDB ID	Resolution (Å)	Components	Organism	Publication
5GAN	3.7	30 proteins + U4/U6 and U5 snRNAs	*S. cerevisiae*	Nguyen, Galej, Bai, et al. (2016)
3JCM	3.8	U5 snRNA, U4/U6 snRNA, 30 proteins Prp8, Snu114, pre-mRNA	*S. cerevisiae*	Wan et al. (2016)
3JB9	3.6	37 Proteins, three snRNAs, 1 RNA lariat	*S. pombe*	Yan et al. (2015)

dissect such a complex, highly dynamic and heterogeneous machine in a holistic manner. Second, splicing is still studied in a crude yeast whole-cell extract (WCE) in vitro, where multistep reversible processes are obscured by asynchronous progression along the pathway. Third, splicing in WCE—developed in 1985 (Brody & Abelson, 1985)—is generally inefficient, leading to only a fraction of molecules undergoing one or both steps of splicing (Fig. 2). In addition, while RNA (–protein) conformational rearrangements often occur at subsecond timescales, it is known that pre-mRNAs are spliced in vitro only over 15–30 min (Das et al., 2006), requiring techniques for in vitro studies that span a wide range of timescales.

Single-molecule fluorescence resonance energy transfer (smFRET) presents a solution uniquely suited to these challenges and can provide dynamic information to animate the emerging static structures, a critically important step toward understanding the splicing mechanism. More specifically, we have developed prism-based total internal reflection fluorescence (P-TIRF) microscopy-based smFRET assays that monitor the distances between site-specifically placed donor and acceptor fluorophores (Figs. 1A, 2, and 3). Alternatively, a wide-field objective-type total internal reflection fluorescence microscope can be used (Roy, Hohng, & Ha, 2008). Analysis of the resulting time trajectories using smFRET histograms, hidden Markov models (HMMs), and various transition density plots (TDPs) has begun to dissect pre-mRNA conformational changes during splicing (Abelson et al., 2010; Blanco, Johnson-Buck, & Walter, 2012; Blanco et al., 2015; Blanco & Walter, 2010; Johnson-Buck, Blanco, & Walter, 2012; Kahlscheuer, Widom, & Walter, 2015; Krishnan et al., 2013; Semlow, Blanco, Walter, & Staley, 2016; Widom, Dhakal, Heinicke, & Walter, 2014). In particular, we identified a 135 nucleotide efficiently spliced yeast pre-mRNA, *UBC4*, in which donor and acceptor were placed either in the exons adjacent to the 5′SS and 3′SS or in the exon and intron adjacent to the 5′SS and BP, giving us unique windows into substrate dynamics. A strength of single-molecule approaches is that the macromolecular machines in the sample need not be synchronized, allowing the observation of both equilibrium and nonequilibrium processes (Walter, Huang, Manzo, & Sobhy, 2012). This is especially important when trying to understand highly reversible conformational steps, which are required in biological systems for proofreading functions and to facilitate pathway bifurcations as encountered in, for example, alternative splicing. However, the sheer number of states and interstate transitions became a challenge early on (Abelson et al., 2010), prompting us to develop smFRET techniques that employ

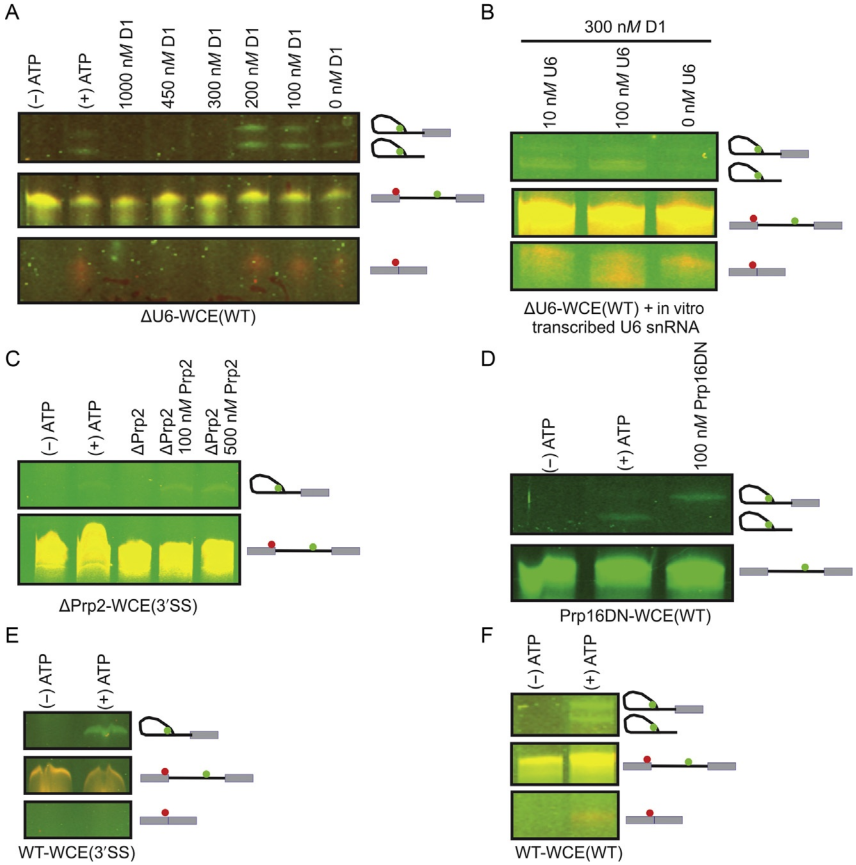

Fig. 2 Confirmation of blockage and reconstitution of splicing by in vitro splicing assays. Denaturing polyacrylamide gels were scanned with a variable mode Typhoon imager. The intron and intron-lariat products are observed in the Cy3 scan (*green*), and the mature mRNA product is visualized in the Cy5 scan (*red*). (A) The optimized concentration of oligodeoxynucleotide D1 required to deplete U6 snRNA (300 n*M*) was determined by titrating increasing amounts of D1 into the in vitro splicing assay. (B) Using the previously determined optimal concentration of D1 (300 n*M*, A), extract viability was confirmed through reconstitution with in vitro transcribed U6 snRNA. (C) Incubation of *prp2-1 cef1-TAP* yeast whole-cell extract at 37°C for 40 min completely blocks splicing activity (ΔPrp2 lane). Addition of recombinant Prp2p to the 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 3′SS mutant substrate with WCE stalls splicing after the first step, while incubation with a WT substrate (F) results in efficient progression through both steps of splicing. *Reproduced with permission from Blanco, M. R., Martin, J. S., Kahlscheuer, M. L., Krishnan, R., Abelson, J., et al. (2015). Single molecule cluster analysis dissects splicing pathway conformational dynamics.* Nature Methods, 12*, 1077–1084.*

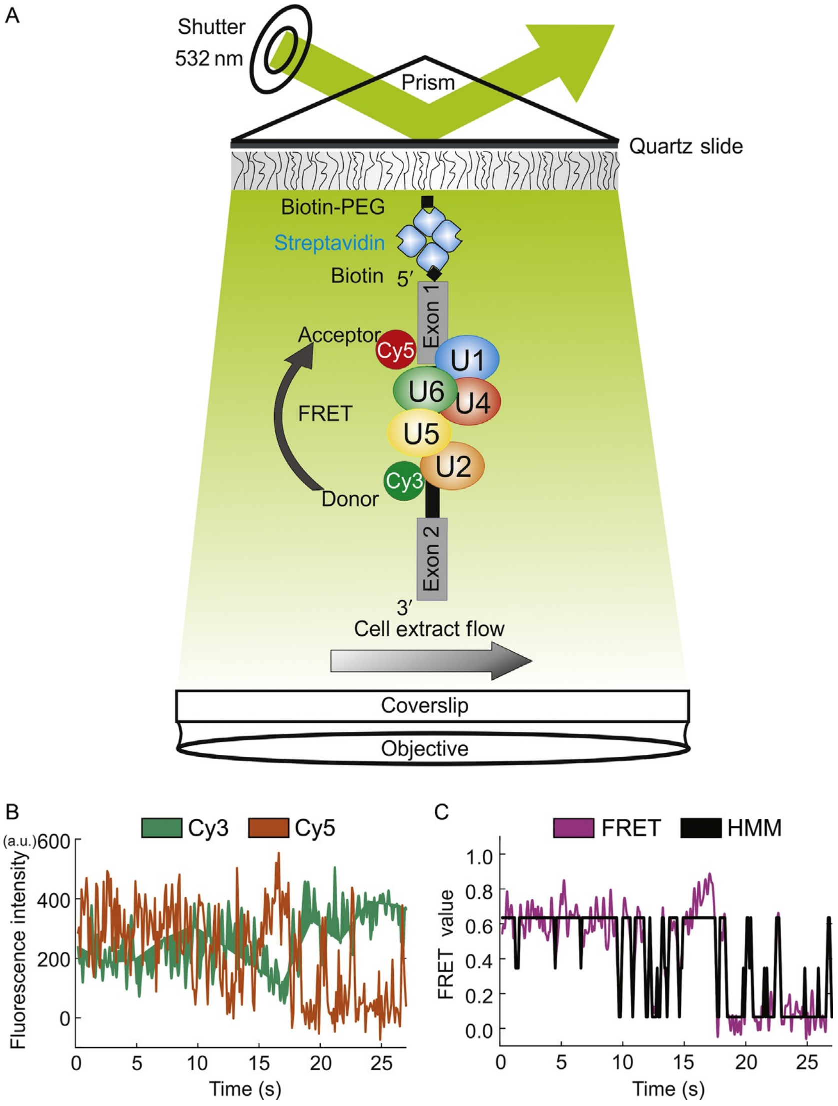

Fig. 3 Typical acquisition of smFRET data. (A) Prism-based TIRF microscope with microfluidic sample chamber containing the 5′ biotinylated Ubc4 pre-mRNA substrate incubated with various yeast splicing extracts (Abelson et al., 2010; Blanco et al., 2015). (B) Raw single-molecule time trace showing the anticorrelated donor (*green*) and acceptor (*red*) intensities. (C) The corresponding FRET trace (*purple*) and the HMM trace as assigned by vbFRET (*black*). *Reproduced in part with permission from Blanco, M. R., Martin, J. S., Kahlscheuer, M. L., Krishnan, R., Abelson, J., et al. (2015). Single molecule cluster analysis dissects splicing pathway conformational dynamics.* Nature Methods, 12, *1077–1084.*

either biochemical enrichment of specific complexes via glycerol gradient centrifugation and immunoprecipitation (termed Single-Molecule Pull-Down FRET, or SiMPull-FRET; Kahlscheuer et al., 2015; Krishnan et al., 2013; Semlow et al., 2016) or bioinformatic clustering for "computational purification" termed Single-Molecule Cluster Analysis, SiMCAn (Blanco et al., 2015) to pinpoint the structural dynamics of specific splicing intermediates. Our approach of focusing on the pre-mRNA substrate with SiMPull-FRET and SiMCAn is particularly timely given the emerging cryo-EM structures of relevant spliceosomal complexes typically lack at least part of the substrate (Fig. 1C and Table 1) and can thus be uniquely complemented and animated by substrate data.

SiMCAn translates time-resolved smFRET data into a temporal sequence of conformational changes that single pre-mRNA molecules adopt throughout splicing. In its initial application, SiMCAn discovered a previously undescribed low-FRET conformation adopted late in splicing by a 3′SS mutant (Blanco et al., 2015). Conversely, the growing number of high-resolution structures of spliceosomal components and complexes will aid in the refinement of SiMCAn-derived conformational models, allowing us to test the hypothesis that specific sequences of structural changes facilitate splicing. As an ultimate goal, SiMCAn has the potential to help delineate the mechanism of pre-mRNA splicing by dissecting—at single-molecule and single-nucleotide resolution—the complete "movie" of conformational changes throughout a full splicing cycle (Fig. 1), or any other biomolecular process. MATLAB scripts to perform all SiMCAn analyses are available as Supplementary Software with our original publication (Blanco et al., 2015).

2. EXPERIMENTAL METHODS AND DATA ANALYSIS

Data from smFRET studies have directly established that populations of single molecules often exhibit variant behaviors that include dominant, rare, long-lived, or transient species, as well as singular and parallel (i.e., alternate) pathways. Observations of large populations of molecules provide a statistical sample that mirrors ensemble measurements, while also being sensitive enough to describe the component subpopulations. The strengths of smFRET methods have established them as a useful probe of RNA molecular catalysis and dynamics, yet their application to more complex and multistep systems faces additional hurdles. Molecular systems that display two or three prominent FRET states are often accurately identified using either thresholding or HMM approaches (Blanco & Walter, 2010).

However, the analysis of multistate smFRET data may be subject to user judgment and bin size selection bias. With a higher number of states, the analysis becomes increasingly unreliable. Spliceosomally catalyzed pre-mRNA splicing in particular is composed of a multitude of conformational states and substates that rapidly interconvert. Analysis of smFRET trajectories from pre-mRNA splicing in vitro required new approaches to ensure unbiased sampling and accurate modeling of the interactions and rearrangements that take place between the RNA, protein, and pre-mRNA components of the spliceosomal machinery. Understanding the subtlety of these dynamics required comparison of smFRET trajectories featuring different conformational states and kinetic properties from across a range of experimental conditions. The complexity of pre-mRNA splicing precluded fitting all data with a single HMM model, which would impose a constraining single, preordained kinetic model on all molecules and conditions.

More generally, the results from kinetic measurements at the single-molecule level are fundamentally different from classical ensemble-averaged data due to their statistical nature (Moffitt, Chemla, & Bustamante, 2010). When we began addressing the challenge of mining the increasingly complex data obtained from spliceosomal complexes for commonalities, we realized that a simple alignment of time trajectories did not yield satisfactory results since the dwell time in any given state is part of an exponential probability distribution. That is, single molecules with both very long and very short dwell times in a given state are found regardless of the underlying rate constants, which only modulates the probability of their observation (and short dwell times are always the most likely). We therefore returned to using HMMs, aiming to overcome their underlying limitations. In particular, a Markov transition from state A to state B is assumed to take place with total state memory loss, a feature that is fundamentally not the case for a biological process wherein a specific functional intermediate is a product of all preceding steps and itself funnels the biomolecular machine into a subset of accessible subsequent steps. Put differently, all of biology is inherently dependent on energy being used to bias the direction of molecular interactions and transformations, thus overcoming the dominance of complete randomness dictated by the Second Law of Thermodynamics. In the case of biomolecular machines such as the spliceosome that operate at the nanoscale—where inertia and friction have lost their preponderant role exerted at the macroscale—intrinsic, random thermal motions are biased directionally by helicases that, as molecular motors, utilize the energy from ATP hydrolysis to achieve biased Brownian ratcheting (Kahlscheuer et al., 2015;

Krishnan et al., 2013). To discern this nonrandomness in our smFRET data, we turned to a hybrid method wherein HMMs are used to capture and summarize the molecular FRET states of single molecules and the transitions between them in a FRET similarity matrix (FSM), followed by "alignment" of the single-molecule FSMs to identify clusters of common molecular behaviors (Blanco et al., 2015). A second level of clustering based on the abundance of dynamic behaviors when splicing is stalled at defined intermediate steps through biochemical and genetic tools then allowed us to assign pre-mRNA FRET states and transitions to specific splicing complex intermediates. In the following, we elaborate on relevant considerations for this type of analysis.

2.1 Acquisition and Initial Analysis of smFRET Data

Clustering analysis has been used in a variety of fields to help group large datasets (e.g., gene expression profiles) based on a similarity metric that distinguishes similar from disparate behaviors (D'haeseleer, 2005). Computational clustering provides the necessary analytical power to tease out individual subpopulations from a larger single-molecule dataset. We here discuss the application of SiMCAn (Blanco et al., 2015) a hierarchical clustering tool to extract the consensus behaviors of pre-mRNAs throughout the assembly and catalytic steps of the yeast spliceosome. The technique is generalizable for other datasets with more or less complicated dynamics whose analysis by traditional smFRET tools is insufficient, too error prone or too cumbersome to extract the full scale of information available. The algorithms to perform all SiMCAn analyses are available as MATLAB scripts (The MathWorks, Inc., Natick, MA) as part of the Supplementary Software of our original publication (Blanco et al., 2015).

In smFRET, a donor fluorophore (often Cy3) is excited with typically a 532-nm laser, with the resulting emission of both donor and acceptor fluorophore (often Cy5) recorded on a P-TIRF microscope setup (Fig. 3A). Importantly, SiMCAn is truly agnostic to the nature and origin of the data. Molecules selected for further analysis by SiMCAn should fulfill certain criteria, however. The molecule's fluorescence time trajectory should last at minimum, ~30 data points (or, more generally, at least three times the mean dwell time of the observed FRET states) before photobleaching of the donor fluorophore (Cy3), show anticorrelated changes in donor (Cy3) and acceptor (Cy5) intensity, undergo single-step photobleaching, and contain active acceptor (Cy5) fluorophore throughout, as evidenced upon direct 635-nm laser excitation at the end of each data acquisition (Blanco et al., 2015). After extracting the signal, FRET values are calculated by dividing the cross talk corrected intensity of the acceptor emission by the total emission from both donor and acceptor (Fig. 3B and C). Trajectories are truncated to just before the first fluorophore photobleaches. To reliably identify genuine FRET states and their transition kinetics in inherently noisy datasets, each selected smFRET trace is then fitted with an HMM, which traces the most probable sequence of FRET states and derives their dwell times, using vbFRET (Bronson, Fei, Hofman, Gonzalez, & Wiggins, 2009) in the MATLAB environment with no assumptions about the values or distributions (Fig. 3C). In principle, any idealized trajectories from HMM-fitted data (generated using vbFRET, HaMMy, QuB, etc.) (Blanco & Walter, 2010) subsequently can be fed into SiMCAn. Multiple models with differing number of states should initially be used to determine the underlying FRET states; typically, the entire dataset for each condition is analyzed by the iterative application of the Viterbi and BaumWelch algorithms to generate idealized trajectories (Blanco & Walter, 2010). The number of states assumed in the idealization can be varied from 2 to 11 and the corresponding fits evaluated using the Bayesian information criterion (BIC). The number of states that result in the best BIC score should be used in the subsequent analysis. Using the BIC penalizes models with extraneous states that do not result in a significant improvement in the LogLikelihood, allowing one to select the most appropriate model by balancing goodness of fit and model parsimony. Many of the currently available software packages for HMM fitting of smFRET trajectories have this procedure incorporated into their analysis.

For pre-mRNA splicing, we binned each HMM assigned FRET state into the closest of 10 evenly spaced FRET values (0.05–0.95, increment of 0.10), chosen in relation to the error of ±0.05 expected from a measurement of a standard FRET value in the range of 0 and 1 (Abelson et al., 2010; Blanco et al., 2015). This information is already much more fine-grained than a simple FRET occupancy histogram, which plots frame-by-frame FRET values for a set of molecules collected over the first 100 frames (10 s of real time) of all FRET trajectories in a given dataset, equivalent to typical ensemble FRET data (Fig. 4A; Blanco & Walter, 2010). In such a histogram, all information on the connectivity of specific states, their interconversion kinetics, and variance in FRET value is lost. A more appropriate way to capture the details of HMM data of many molecules is the use of TDPs, weighted by the number of times a particular FRET transition

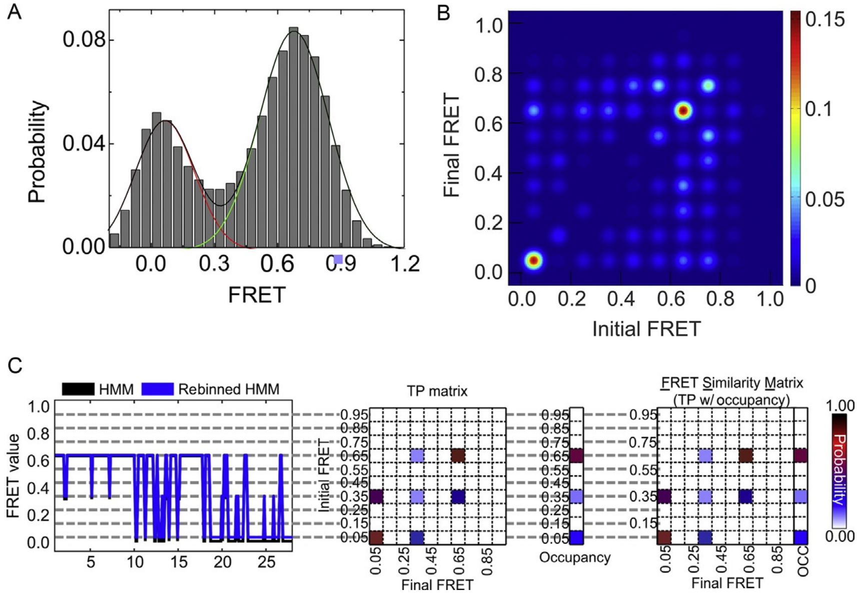

Fig. 4 Basic smFRET data analysis. (A) FRET probability distribution, summarizing all FRET values observed over the first 100 data points of 6079 molecules. (B) Transition occupancy density plot (TODP) for the data shown in (A), represented as a heat map of the fraction of molecules showing any particular transition. (C) Assigned FRET trace before (*black*) and after (*blue*) reassignment to the closest of 10 evenly spaced states (0.05–0.95, increments of 0.10, *gray dashed lines*). The transition probability (TP) matrix corresponds to the rebinned FRET trace in the *left panel* and represents the transition probability between each of the 10 binned FRET values. The FRET similarity matrix (FSM) combines both the TP matrix and the FRET state occupancies over the entire smFRET trace. *Reproduced in part with permission from Blanco, M. R., Martin, J. S., Kahlscheuer, M. L., Krishnan, R., Abelson, J., et al. (2015). Single molecule cluster analysis dissects splicing pathway conformational dynamics.* Nature Methods, 12, *1077–1084.*

occurs, or transition occupancy density plots (TODPs), weighted by the fraction of molecules displaying a particular transition (Figs. 4 and 5; Blanco & Walter, 2010). Of note, to eliminate possible effects of the high concentrations of proteins directly on the fluorophores, one may additionally evaluate the separated donor and acceptor signals by HMM and introduce a scoring function to select for anticorrelated donor and acceptor changes that reflect verified changes in FRET (Blanco & Walter, 2010). To better parse the complexity of pre-mRNA splicing, we decided to enrich specific splicing intermediates through seven biochemical and genetic stalls or combinations thereof (Fig. 2), as well as recorded smFRET data within

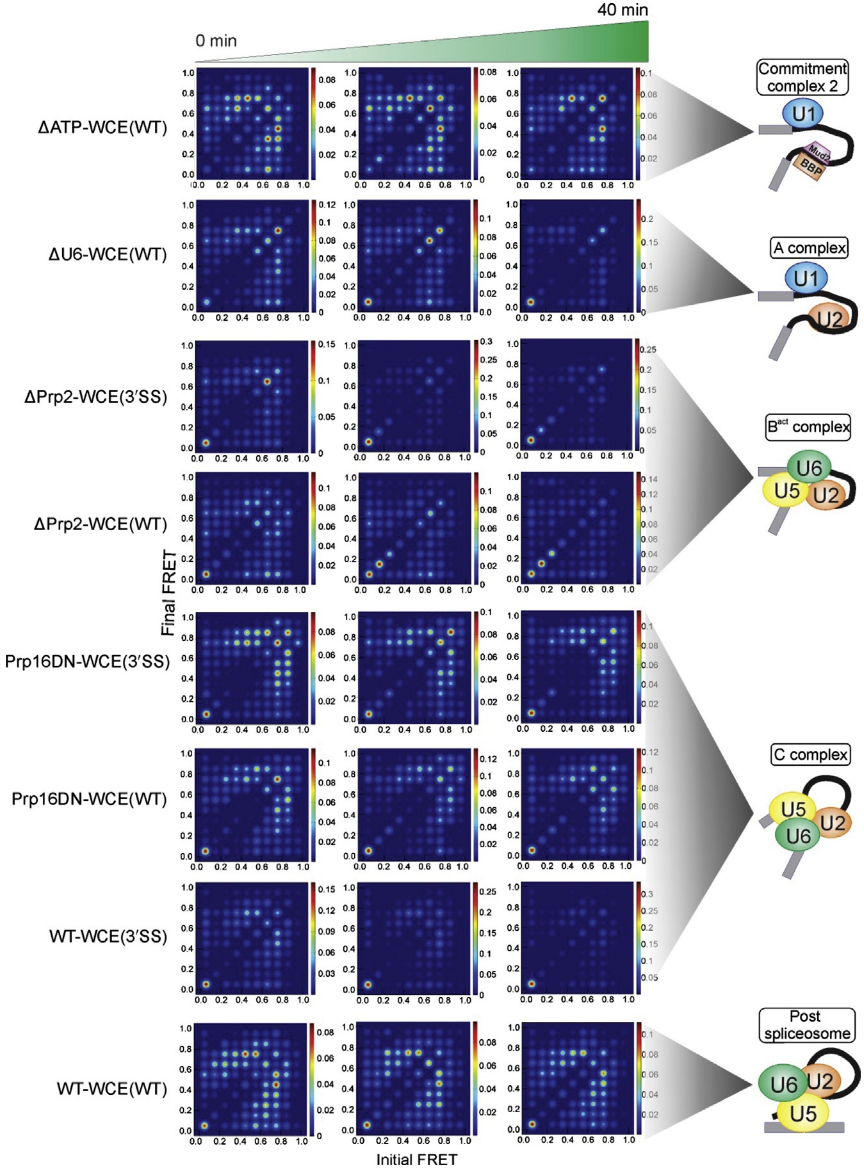

Fig. 5 Transition occupancy density plots (TODPs) from smFRET-monitored pre-mRNA splicing experiments. For each of the eight experimental conditions over the time course of the smFRET experiments, TODPs were used to depict the most probable transitions between an initial FRET state (x-axis) and a final FRET state (y-axis). *Reproduced with permission from Blanco, M. R., Martin, J. S., Kahlscheuer, M. L., Krishnan, R., Abelson, J., et al. (2015). Single molecule cluster analysis dissects splicing pathway conformational dynamics.* Nature Methods, 12*, 1077–1084.*

three time windows of 0–8 min (early), 18–23 min (middle), and 33–40 min (late) after starting each assay. Including the absence of any stall, we collected a total of 10,680 time traces and corresponding HMMs (Blanco et al., 2015). The resulting TODPs show significant differences, but commonalities across conditions are challenging to discern by eye (Fig. 5), necessitating a more sophisticated analysis as offered by SiMCAn.

2.2 SiMCAn: Representing Single Molecules as FRET Similarity Matrices

In SiMCAn, each HMM of a single molecule is used to construct a symmetric transition probability (TP) matrix that delineates the molecule's FRET states as well as the probability per time unit for transitioning from one to any other state, including itself (Fig. 4C). The TP matrix thus summarizes for each molecule all conformational states it adopts and their interconversion kinetics, at a level of precision commensurate with the error of the underlying smFRET measurement, in our case as a 10 × 10 matrix (Blanco et al., 2015). Each TP matrix is then combined with an additional column describing the percent of time (before photobleaching) a given trace spends in each FRET state, resulting in the FSM (Fig. 4C). SiMCAn thus reduces each single-molecule trajectory to an easily comparable unit of information, the FSM, a data abstraction that is possible regardless of the number of FRET states or distinct molecular behaviors among the sampled species. The FSM allows the user to compare the multitude of FRET trajectories through hierarchical cluster analysis in an unbiased fashion, as described in the following section.

2.3 SiMCAn: The First Layer of Clustering

The FSM datasets collected for all experimental conditions are mixed, and then a pairwise distance between all possible trajectories is calculated using a Euclidean distance metric. In the case of pre-mRNA splicing, prior to SiMCAn clustering 4601 "static" molecules (no transitions with occupancy in only one state) were identified based on their FSMs among our total of 10,680 time traces and analyzed separately. Their features are simple so that they can be treated separately, and they are removed to avoid that they dominate clusters of other dynamic molecules with long lived, slowly interconverting FRET states. It cannot be distinguished whether these "static" molecules truly reside in only one state, adopt one long-lived state that happened to photobleach before transitioning into another state, or instead reside in two or more states so short lived that they time average into what only appears, at the available time resolution, as a single state with a mean FRET value.

The remainder, in our case the FSMs of 6079 dynamic molecules, are clustered using Ward's method for hierarchical clustering, in which the Euclidean distance between any pair of clusters r and s is given by Eq. (1):

$$d(r,s)=\sqrt{\frac{2n_rn_s}{n_r+n_s}}\left(\bar{x}_r^2-\bar{x}_s^2\right)^{1/2}, \tag{1}$$

where n is the number of elements and x is the centroids for the respective clusters r and s. This clustering algorithm is implemented natively in MATLAB. The goal is to group together molecules with short distances (high similarity) and connect them with other groups of molecules based on their average intercluster distance. This hierarchical clustering produces a large tree where each molecule is represented by a branch end point, or leaf (Fig. 6A). This tree can be cut (pruned) at various heights, leaving more or less clusters. A pruning threshold needs to be selected so that the intracluster distance (similarity within a cluster) is minimized and the intercluster distance (dissimilarity from other clusters) is maximized. The cutoff for this ideal number of clusters to represent a given dataset is best determined by Eq. (2), a k-means calculation as follows:

$$k=\frac{1}{B}\sum\left(\log\left(W_k^r\right)-\log\left(W_k\right)\right), \tag{2}$$

where B is the number of cluster sets and W_k^r is the average intercluster distance in a random set of k clusters, while W_k is the average intercluster distance of the actual clusters. In the case of pre-mRNA splicing, we pruned the tree to a height of 25 distinct clusters based on an inflection point encountered in the k-means plot over the number of clusters (Fig. 6B). That is, while the k-means used to evaluate the intercluster distances decreases monotonically as more clusters are chosen, at some point the number of clusters is so high that many similar FRET traces become separated. One therefore judiciously selects an inflection point in the k-means plot where an additional increase in the number of clusters has a smaller effect than the last increase had. The goal is that each final cluster represents a unique, distinguishable molecular behavior. For our splicing data, the resulting 25 dynamic clusters were named 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) and were combined with the 10 static clusters, named for their sole FRET state (e.g., 0.05-S). Bootstrap analysis can be used to ask whether appropriate clusters capture all additional FSMs added to the

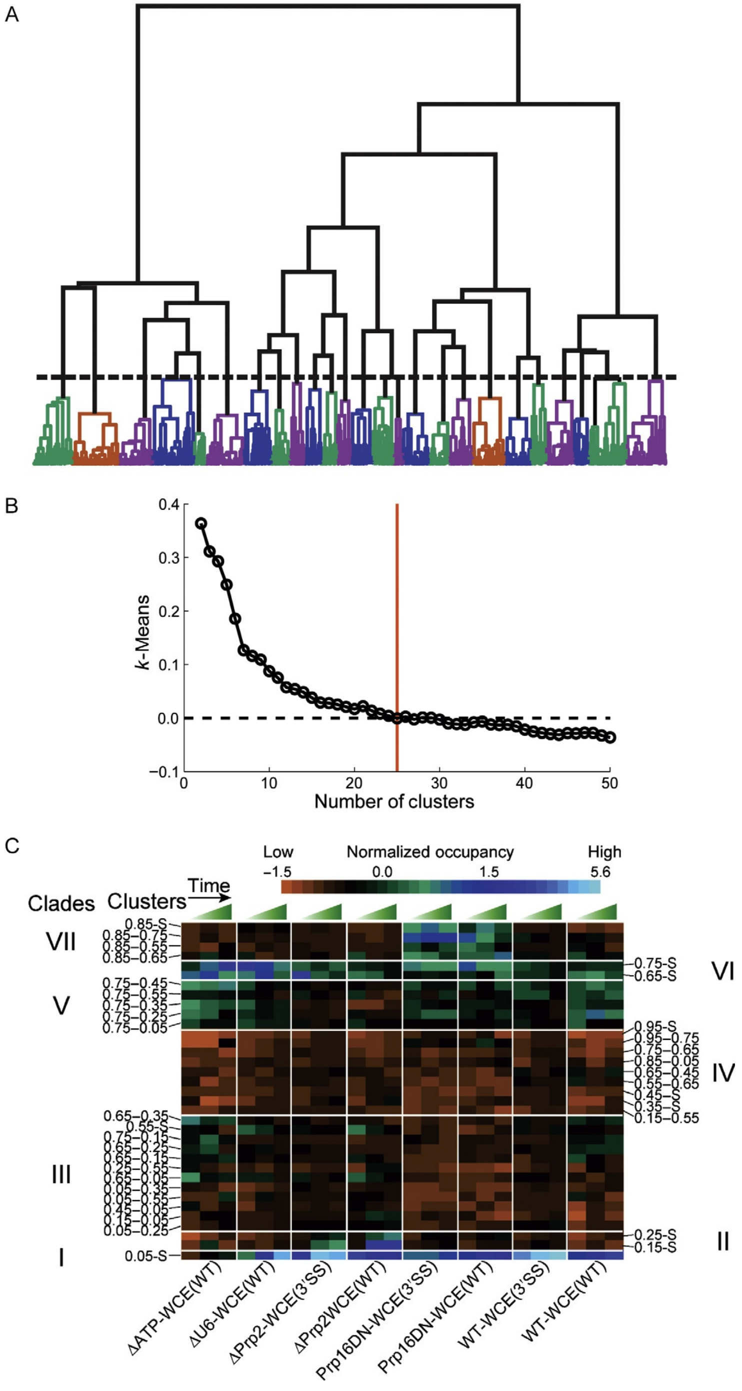

Fig. 6 See figure legend on next page.

dataset. This first layer of SiMCAn clustering is performed sufficiently unsupervised to generate a model-free tree to represent the types of molecular behaviors encountered in a given dataset objectively, without constraining the analysis with any a priori expectations.

2.4 SiMCAn: A Second Layer of Clustering Links the Data to Biology

In the case of pre-mRNA splicing, substrate mutations or extract modifications that block splicing at specific intermediates (Fig. 1B) can be used to limit the observed smFRET dynamics to steps prior to that block, thus enriching certain molecular behaviors and limiting others. This combination of a first, model-free round of clustering with a second round incorporating a priori knowledge from biochemically and genetically manipulated yeast extracts allowed further dissection of splicing in yeast. In particular, we sought to identify clusters whose occupancies are similarly enriched or depleted for a group of conditions, i.e., follow a similar pattern of high and low occupancies across these conditions, suggesting that they can be grouped into a "clade" (or cluster by a different name) of clusters (Fig. 6C). To this end, the occupancy within all 35 first-round clusters across

Fig. 6 SiMCAn-based cluster analysis. (A) Hierarchical clustering of 6079 dynamic single molecules. Varying the tree cutoff heights upon grouping the cluster occupancy among the eight experimental conditions leads to distinct numbers of (color-coded) clades of clusters. smFRET data were collected for each substrate and extract mutant and globally fit to a 10-state HMM. An FSM was then constructed for each molecule and clustered. Each colored branch represents a set of molecules that share common FRET transition probabilities. The *dashed line* indicates the threshold of 25 clusters used to describe the data. Molecules with only one state (static) were clustered separately and later incorporated as separate clusters. Hierarchical clustering was performed utilizing MATLAB routines by calculating the pairwise Euclidean distance of every single molecule and building an agglomerative hierarchical cluster tree utilizing the unweighted average distance to compute distances between clusters. (B) 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. (C) A heat-map representation of the clustering of clusters for the eight experimental conditions. Clusters are group based on profiles to identify "clades" of similar abundance patterns. The abundance profiles for each cluster were constructed and clustered to determine if groups of clusters were shifting in groups, similar to groups of genes with similar expression profiles. This procedure was performed using MATLAB's clustergram function. *Reproduced in part with permission from Blanco, M. R., Martin, J. S., Kahlscheuer, M. L., Krishnan, R., Abelson, J., et al. (2015). Single molecule cluster analysis dissects splicing pathway conformational dynamics.* Nature Methods, 12, *1077–1084.*

all conditions was used as a new similarity matrix to compute their distances, again based on a Euclidean distance measurement. Clades were generated by the same iterative *k*-means approach as before, with the aim to generate a reasonable group of clusters whose occupancy patterns across conditions are most alike (as measured by the Euclidean distance). For our splicing dataset, this second level of clustering yielded an optimal tree height of seven clades that was able to capture the changes in dynamic behavior as the pre-mRNA progresses through the splicing cycle. For example, conditions that stall WT substrate splicing at the early A complex (i.e., ΔU6-WCE(WT), Fig. 1B) show a time-dependent increase in clade I (Fig. 6C; Blanco et al., 2015). This stable low-FRET conformation suggests that the pre-mRNA substrate is partially denatured in the A complex compared to pre-mRNA in the absence of spliceosomal components (Abelson et al., 2010), keeping the first-step reaction sites at a distance. As another example, under the ΔPrp2-WCE(WT) and ΔPrp2-WCE(3′SS) conditions, which both should similarly enrich for the activated B^{act} spliceosome before the first step of splicing since the 3′SS mutation only affects the second step (Fig. 1A and B), SiMCAn recognizes a pair of dominant low-FRET clusters, 0.25-S and 0.15-S, that are appropriately grouped into clade II (Fig. 6C). Since these two clusters are so close in FRET value, they likely represent the same static low-FRET B^{act} conformation found independently upon biochemical purification of the B^{act} complex (Krishnan et al., 2013). Notably, SiMCAn was able to distinguish these clusters from the similarly static, but even lower FRET cluster 0.05-S of the A complex, which is not resolvable in an smFRET histogram (Fig. 4A).

In this way, the second clustering based on SiMCAn can reveal known or anticipated features of specific spliceosomal intermediates. Significantly, SiMCAn also identified heretofore undescribed differences in molecular behavior between the WT and 3′SS mutant substrates upon incubation with WT splicing extract, i.e., comparing conditions WT-WCE(WT) and WT-WCE(3′SS). In particular, both substrates are expected to progress through most of the splicing cycle until the second chemical step (Fig. 1B) and accordingly exhibited quite a similar set of pre-mRNA conformations (Fig. 6C). However, the 3′SS over time additionally adopts a static 0.05-S cluster, suggesting that the 5′SS and BP are far separated in comparison to their relative proximity in the splicing intermediate prior to this point in the cycle (Fig. 1B), represented by conditions Prp16DN-WCE(WT) and Prp16DN-WCE(3′SS) (Fig. 6C). This 0.05-S state is more predominant for the 3′SS mutant than the WT substrate, consistent with 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 first detected (Blanco et al., 2015). This discovery of a proofreading step, which allows for the selection of alternative 3′ splice sites if the canonical 3′SS is removed or damaged through mutation, was subsequently confirmed by smFRET probing of biochemically isolated 3′SS mutant complexes (Semlow et al., 2016), demonstrating the power of SiMCAn in discovery.

2.5 Generalized Application of SiMCAn to Datasets of Varying Complexity

Generally, experimental sampling ought to scale with the total number of distinguishable FRET states and kinetics in a dataset to ensure a reasonable level of convergence of SiMCAn, as with any other analysis approach. We developed SiMCAn specifically for dealing with very large datasets of high complexity, where traditional HMM analysis becomes inadequate and commonalities between single-molecule traces are challenging to identify. The availability of a dataset with >10,000 smFRET traces as described earlier for the pre-mRNA splicing cycle is, however, still not typical in the field. An example of a dataset more typically encountered in current single-molecule studies is that of the Prp2-mediated conformational transition immediately prior to the first step of splicing (Krishnan et al., 2013). In short, the immobilized B^{act} complex containing FRET-labeled pre-mRNA was monitored as it progresses through the B* to the C complex upon addition of recombinant proteins Prp2, Spp2, and Cwc25 (Fig. 7A). Notably, SiMCAn rapidly (within minutes) identified subpopulations of pre-mRNA molecules, recapitulating the subpopulations of molecules that had originally been identified manually by visual inspection (Fig. 7; Krishnan et al., 2013). More specifically, the HMM-fitted FRET traces under conditions designed to enrich for the B^{act}, B*, and C complexes, respectively, were converted into FSMs and analyzed together using SiMCAn. Again, maximizing the intercluster distances while minimizing the intracluster distances using SiMCAn revealed nine dynamic and four static clusters as most consistent with the data (Fig. 7B). As seen in bar graphs of the fraction of molecules present in each cluster under each of the three experimental conditions (Fig. 7C), we found that a cluster of static low-FRET (0.3-S) molecules dominates the B^{act} conditions, consistent with our previous observation of such a dominant state in the B^{act} spliceosome (Blanco et al., 2015; Krishnan et al., 2013). By contrast, a static high-FRET cluster (0.7-S) is most

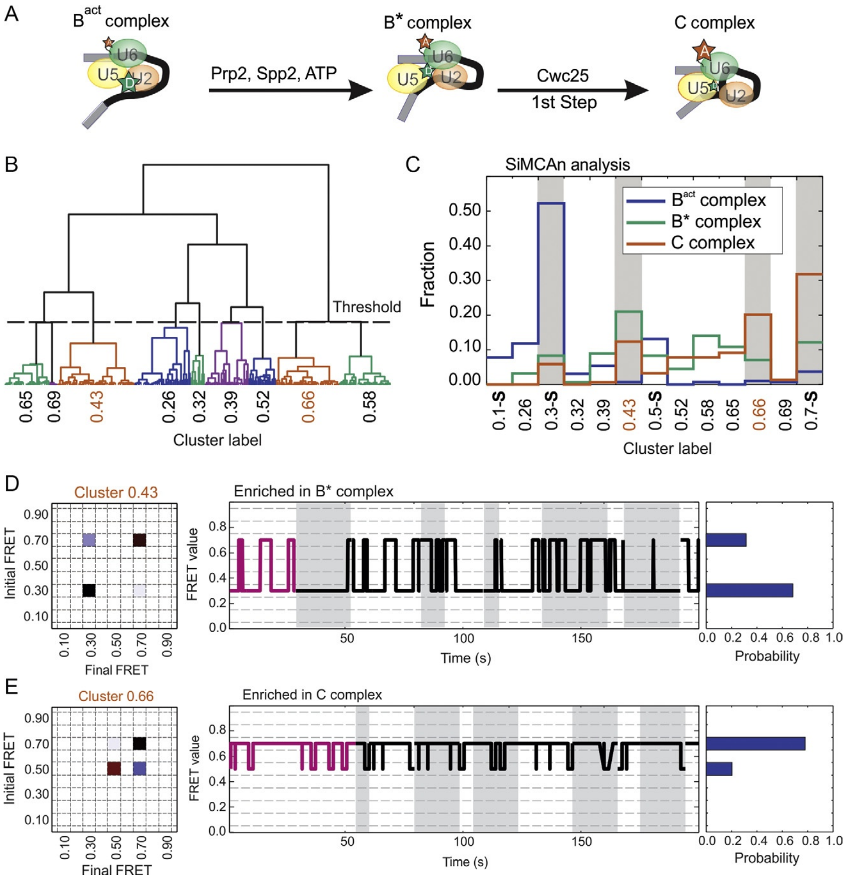

Fig. 7 Performance of SiMCAn on a previously analyzed smaller dataset describing the transition from the purified B^{act} to the C complex (Krishnan et al., 2013). (A) Protein requirements for the transition from the B^{act} complex through B* to the C complex. (B) Hierarchical tree based on hierarchical clustering analysis of the dynamic molecules refits with FRET states of 0.1, 0.3, 0.5, and 0.7 (Krishnan et al., 2013). Static molecules were identified and analyzed by SiMCAn separately. (C) Cluster occupancy bar graph showing the fraction of molecules from each experimental condition that occupy the nine dynamic and four 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). *Gray bars* highlight the most populated clusters occupied by each of the complexes. (D and E) Dynamic clusters enriched in the B* (D, cluster 0.4267) and C (E, cluster 0.6478) complexes. Each representative for the B* and C complex 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*). *Reproduced with permission from Blanco, M. R., Martin, J. S., Kahlscheuer, M. L., Krishnan, R., Abelson, J., et al. (2015). Single molecule cluster analysis dissects splicing pathway conformational dynamics.* Nature Methods, 12, *1077–1084.*

abundant in the C complex (Fig. 7C), again consistent with manual sorting (Krishnan et al., 2013). Finally, SiMCAn also identified two dynamic clusters increasingly populated under B* (cluster 0.43, green) and C (cluster 0.66, red) complex conditions (Fig. 7C). Cluster 0.43 contains molecules with a short-lived high-FRET state most abundant under B* conditions, whereas cluster 0.66 contains molecules with a longer-lived high-FRET state that is enriched once Cwc25 has been added to favor the C complex (Fig. 7C), together matching our previous manual analysis (Krishnan et al., 2013). These results demonstrate that, when applied to a more typical experimental dataset, SiMCAn is able to segregate the data efficiently based on FRET states and differences in state-to-state interconversion kinetics in a way that derives biologically meaningful results, saves time and avoids potential user bias.

3. CONCLUSIONS AND OUTLOOK

The complex dynamic behavior of the many individual molecules we observed undergoing the pre-mRNA splicing cycle of budding yeast in vitro required the development of a novel approach to smFRET analysis. In recent years, hidden Markov modeling has vastly improved the amount of information one is able to extract from smFRET data. It is a powerful tool for the unbiased extraction of FRET states and kinetics from single-molecule trajectories. To leverage the TP matrix resulting from hidden Markov modeling, we created an FSM that can be utilized in hierarchical clustering algorithms. In turn, hierarchical clustering techniques allow us to group molecules with similar behaviors across conditions by applying the FSM as a measure of how similar any two molecules are. The FSM can help distinguish molecules with similar FRET states and rates of interconversion using SiMCAn, a tool available as supplemental MATLAB scripts associated with our original publication (Blanco et al., 2015). As a first application, we analyzed single-molecule data from pre-mRNAs labeled at the 5′SS and BP that were exposed to various yeast WCEs competent in splicing. To enrich for specific points along the splicing cycle, we introduced substrate mutations and depleted specific splicing factors. Blocking specific points along the splicing pathway allowed us to focus on dynamics that are relevant during the early and late stages of assembly, identifying both known and heretofore undescribed conformational states.

The ultimate goal in studies of any enzyme is to understand the link between the temporal sequence of conformational fluctuations and the chemical transformation(s) throughout a full catalytic cycle; in fact, it can be argued that the former begets the latter (Hammes-Schiffer & Benkovic, 2006). So far, this goal has been accomplished mostly for single-chain metabolic enzymes such as dihydrofolate reductase (Bhabha et al., 2011; Hammes-Schiffer & Benkovic, 2006) and adenylate kinase (Adk) (Henzler-Wildman et al., 2007). The overarching hypothesis underlying our work is that, similarly, a characteristic, helicase-directed sequence of stochastic conformational fluctuations leads to chemical catalysis in the more complex enzymatic machinery of the spliceosome. The power of SiMCAn-based clustering analysis lies in its ability to compare datasets efficiently and objectively by sorting single-molecule time traces with their stochastic idiosyncrasies into clusters of common FRET state and kinetic behaviors, and assessing the relative abundance of each cluster across any number of experimental conditions. SiMCAn thus emerges as a computational approach capable of translating smFRET data into temporal sequences, or "movies," of most common RNA conformational paths through a catalytic cycle that, combined with high-resolution cryo-EM and crystal structures, footprinting and other solution data will eventually make the spliceosome as well understood as single-chain enzymes are today.

Previous work has used clustering algorithms to identify distinct kinetic behaviors, but focused on small datasets with two or three FRET states and limited dynamics (Greenfeld, Pavlichin, Mabuchi, & Herschlag, 2012; Keller, Kobitski, Jaschke, Nienhaus, & Noe, 2014). SiMCAn performs well with such datasets, as demonstrated when applied to the B^{act} to B* to C complex transition, but is adept at handling also more complex systems with higher numbers of states and complex kinetic networks under non-equilibrium conditions. That opens the door to more sophisticated experiments, for example, by labeling additional spliceosomal RNA and protein components. The number of potential targets for single-molecule experiments is great and facilitated by the wealth of genetic and biochemical tools available in the budding yeast splicing system (Hoskins et al., 2011). As more high-resolution structures of the spliceosome become available, they will provide further targets for specific interactions to monitor. The work presented here therefore presents a framework for further advances in what is sure to expand into the exciting and fruitful field of single-molecule splicing. Finally, SiMCAn stands ready to be applied to complex smFRET datasets that are sure to become more common as we expand the use of single-molecule tools to many more biomolecular machines.

ACKNOWLEDGMENTS

This work was supported by NIH Grant GM098023 to N.G.W. We thank Dr. Paul Lund for helpful insights.

REFERENCES

Abelson, J., Blanco, M., Ditzler, M. A., Fuller, F., Aravamudhan, P., et al. (2010). Conformational dynamics of single pre-mRNA molecules during in vitro splicing. *Nature Structural and Molecular Biology*, *17*, 504–512.

Agafonov, D. E., Kastner, B., Dybkov, O., Hofele, R. V., Liu, W. T., et al. (2016). Molecular architecture of the human U4/U6.U5 tri-snRNP. *Science*, *351*, 1416–1420.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., et al. (2010). Deciphering the splicing code. *Nature*, *465*, 53–59.

Bhabha, G., Lee, J., Ekiert, D. C., Gam, J., Wilson, I. A., et al. (2011). A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science*, *332*, 234–238.

Blanco, M. R., Johnson-Buck, A. E., & Walter, N. G. (2012). Hidden Markov modeling. In G. C. K. Roberts (Ed.), *Encyclopedia of biophysics* (pp. 971–975). Springer.

Blanco, M. R., Martin, J. S., Kahlscheuer, M. L., Krishnan, R., Abelson, J., et al. (2015). Single molecule cluster analysis dissects splicing pathway conformational dynamics. *Nature Methods*, *12*, 1077–1084.

Blanco, M., & Walter, N. G. (2010). Analysis of complex single-molecule FRET time trajectories. *Methods in Enzymology*, *472*, 153–178.

Braunschweig, U., Gueroussov, S., Plocik, A. M., Graveley, B. R., & Blencowe, B. J. (2013). Dynamic integration of splicing within gene regulatory pathways. *Cell*, *152*, 1252–1269.

Brody, E., & Abelson, J. (1985). The "spliceosome": Yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. *Science*, *228*, 963–967.

Bronson, J. E., Fei, J., Hofman, J. M., Gonzalez, R. L., & Wiggins, C. H. (2009). Learning rates and states from biophysical time series: A Bayesian approach to model selection and single-molecule FRET data. *Biophysical Journal*, *97*, 3196–3205.

Burgess, S. M., & Guthrie, C. (1993). A mechanism to enhance mRNA splicing fidelity: The RNA-dependent ATPase Prp16 governs usage of a discard pathway for aberrant lariat intermediates. *Cell*, *73*, 1377–1391.

Calarco, J. A., Zhen, M., & Blencowe, B. J. (2011). Networking in a global world: Establishing functional connections between neural splicing regulators and their target transcripts. *RNA*, *17*, 775–791.

Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature Reviews. Genetics*, *3*, 285–298.

Chen, M., & Manley, J. L. (2009). Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. *Nature Reviews. Molecular Cell Biology*, *10*, 741–754.

Chen, W., & Moore, M. J. (2014). The spliceosome: Disorder and dynamics defined. *Current Opinion in Structural Biology*, *24*, 141–149.

Couto, J. R., Tamm, J., Parker, R., & Guthrie, C. (1987). A trans-acting suppressor restores splicing of a yeast intron with a branch point mutation. *Genes and Development*, *1*, 445–455.

Das, R., Dufu, K., Romney, B., Feldt, M., Elenko, M., et al. (2006). Functional coupling of RNAP II transcription to spliceosome assembly. *Genes and Development*, *20*, 1100–1109.

De, I., Schmitzova, J., & Pena, V. (2016). The organization and contribution of helicases to RNA splicing. *WIREs RNA*, *7*, 259–274.

de Klerk, E., & t Hoen, P. A. (2015). Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing. *Trends in Genetics*, *31*, 128–139.

DeLano, W. L. (2002). *The PyMOL Molecular Graphics System*. Palo Alto, CA: DeLano Scientific. *http://www.pymol.org*.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, *23*, 1499–1501.

Greenfeld, M., Pavlichin, D. S., Mabuchi, H., & Herschlag, D. (2012). Single molecule analysis research tool (SMART): An integrated approach for analyzing single molecule data. *PLoS One*, 7, e30024.

Hammes-Schiffer, S., & Benkovic, S. J. (2006). Relating protein motion to catalysis. *Annual Review of Biochemistry*, *75*, 519–541.

Hang, J., Wan, R., Yan, C., & Shi, Y. (2015). Structural basis of pre-mRNA splicing. *Science*, *349*, 1191–1198.

Henzler-Wildman, K. A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., et al. (2007). Intrinsic motions along an enzymatic reaction trajectory. *Nature*, *450*, 838–844.

Horowitz, D. S. (2011). The splice is right: Guarantors of fidelity in pre-mRNA splicing. *RNA*, *17*, 551–554.

Hoskins, A. A., Friedman, L. J., Gallagher, S. S., Crawford, D. J., Anderson, E. G., et al. (2011). Ordered and dynamic assembly of single spliceosomes. *Science*, *331*, 1289–1295.

Jarmoskaite, I., & Russell, R. (2014). RNA helicase proteins as chaperones and remodelers. *Annual Review of Biochemistry*, *83*, 697–725.

Johnson-Buck, A. E., Blanco, M. R., & Walter, N. G. (2012). Single-molecule fluorescence resonance energy transfer. In G. C. K. Roberts (Ed.), *Encyclopedia of Biophysics* (pp. 2329–2335). Springer.

Kahlscheuer, M. L., Widom, J., & Walter, N. G. (2015). Single-molecule pull-down FRET to dissect the mechanisms of biomolecular machines. *Methods in Enzymology*, *558*, 539–570.

Keller, B. G., Kobitski, A., Jaschke, A., Nienhaus, G. U., & Noe, F. (2014). Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models. *Journal of the American Chemical Society*, *136*, 4534–4543.

Koodathingal, P., Novak, T., Piccirilli, J. A., & Staley, J. P. (2010). The DEAH box ATPases Prp16 and Prp43 cooperate to proofread 5′ splice site cleavage during pre-mRNA splicing. *Molecular Cell*, *39*, 385–395.

Koodathingal, P., & Staley, J. P. (2013). Splicing fidelity: DEAD/H-box ATPases as molecular clocks. *RNA Biology*, *10*, 1073–1079.

Kornblihtt, A. R., Schor, I. E., Allo, M., Dujardin, G., Petrillo, E., et al. (2013). Alternative splicing: A pivotal step between eukaryotic transcription and translation. *Nature Reviews. Molecular Cell Biology*, *14*, 153–165.

Krishnan, R., Blanco, M. R., Kahlscheuer, M. L., Abelson, J., Guthrie, C., et al. (2013). Biased Brownian ratcheting leads to pre-mRNA remodeling and capture prior to first-step splicing. *Nature Structural and Molecular Biology*, *20*, 1450–1457.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., et al. (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, *352*, 600–604.

Matera, A. G., & Wang, Z. (2014). A day in the life of the spliceosome. *Nature Reviews. Molecular Cell Biology*, *15*, 108–121.

Mayas, R. M., Maita, H., Semlow, D. R., & Staley, J. P. (2010). Spliceosome discards intermediates via the DEAH box ATPase Prp43p. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 10020–10025.

Mayas, R. M., Maita, H., & Staley, J. P. (2006). Exon ligation is proofread by the DExD/H-box ATPase Prp22p. *Nature Structural and Molecular Biology*, *13*, 482–490.

Moffitt, J. R., Chemla, Y. R., & Bustamante, C. (2010). Methods in statistical kinetics. *Methods in Enzymology*, *475*, 221–257.

Naftelberg, S., Schor, I. E., Ast, G., & Kornblihtt, A. R. (2015). Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annual Review of Biochemistry*, *84*, 165–198.

Nguyen, T. H., Galej, W. P., Bai, X. C., Oubridge, C., Newman, A. J., et al. (2016a). Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution. *Nature*, *530*, 298–302.

Nguyen, T. H., Galej, W. P., Bai, X. C., Savva, C. G., Newman, A. J., et al. (2015). The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature*, *523*, 47–52.

Nguyen, T. H., Galej, W. P., Fica, S. M., Lin, P. C., Newman, A. J., et al. (2016b). CryoEM structures of two spliceosomal complexes: Starter and dessert at the spliceosome feast. *Current Opinion in Structural Biology*, *36*, 48–57.

Nilsen, T. W. (2003). The spliceosome: The most complex macromolecular machine in the cell? *Bioessays*, *25*, 1147–1149.

Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, *463*, 457–463.

Papasaikas, P., & Valcarcel, J. (2016). The spliceosome: The ultimate RNA chaperone and sculptor. *Trends in Biochemical Sciences*, *41*, 33–45.

Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K., Li, J., & Nagai, K. (2009). Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature*, *458*, 475–480.

Poulos, M. G., Batra, R., Charizanis, K., & Swanson, M. S. (2011). Developments in RNA splicing and disease. *Cold Spring Harbor Perspectives in Biology*, *3*, a000778.

Rino, J., & Carmo-Fonseca, M. (2009). The spliceosome: A self-organized macromolecular machine in the nucleus? *Trends in Cell Biology*, *19*, 375–384.

Roy, R., Hohng, S., & Ha, T. (2008). A practical guide to single-molecule FRET. *Nature Methods*, *5*, 507–516.

Scotti, M. M., & Swanson, M. S. (2016). RNA mis-splicing in disease. *Nature Reviews. Genetics*, *17*, 19–32.

Semlow, D. R., Blanco, M. R., Walter, N. G., & Staley, J. P. (2016). Spliceosomal DEAH-Box ATPases remodel pre-mRNA to activate alternative splice sites. *Cell*, *164*, 985–998.

Smith, D. J., & Konarska, M. M. (2008). Mechanistic insights from reversible splicing catalysis. *RNA*, *14*, 1975–1978.

Staley, J. P., & Guthrie, C. (1998). Mechanical devices of the spliceosome: Motors, clocks, springs, and things. *Cell*, *92*, 315–326.

Tseng, C. K., & Cheng, S. C. (2008). Both catalytic steps of nuclear pre-mRNA splicing are reversible. *Science*, *320*, 1782–1784.

Wahl, M. C., Will, C. L., & Luhrmann, R. (2009). The spliceosome: Design principles of a dynamic RNP machine. *Cell*, *136*, 701–718.

Walter, N. G., Huang, C., Manzo, A. J., & Sobhy, M. A. (2012). Do-it-yourself guide: How to use the modern single molecule toolkit. *Nature Methods*, *5*, 475–489.

Wan, R., Yan, C., Bai, R., Wang, L., Huang, M., et al. (2016). The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science*, *351*, 466–475.

Wang, G. S., & Cooper, T. A. (2007). Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nature Reviews. Genetics*, *8*, 749–761.

Widom, J. R., Dhakal, S., Heinicke, L. A., & Walter, N. G. (2014). Single-molecule tools for enzymology, structural biology, systems biology and nanotechnology: An update. *Archives of Toxicology*, *88*, 1965–1985.

Will, C. L., & Luhrmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, *3,* pii: a003707.

Xu, Y. Z., & Query, C. C. (2007). Competition between the ATPase Prp5 and branch region-U2 snRNA pairing modulates the fidelity of spliceosome assembly. *Molecular Cell*, *28*, 838–849.

Yan, C., Hang, J., Wan, R., Huang, M., Wong, C. C., et al. (2015). Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science*, *349*, 1182–1191.

Zhang, J., & Manley, J. L. (2013). Misregulation of pre-mRNA alternative splicing in cancer. *Cancer Discovery*, *3*, 1228–1237.