Single Molecule Cluster Analysis dissects splicing pathway conformational dynamics

Mario R Blanco[1,2,5,6], Joshua S Martin[3,5,6], Matthew L Kahlscheuer[1,6], Ramya Krishnan[1], John Abelson[4], Alain Laederach[3] & Nils G Walter[1]

We report Single Molecule Cluster Analysis (SiMCAn), which utilizes hierarchical clustering of hidden Markov modeling–fitted single-molecule fluorescence resonance energy transfer (smFRET) trajectories to dissect the complex conformational dynamics of biomolecular machines. We used this method to study the conformational dynamics of a precursor mRNA during the splicing cycle as carried out by the spliceosome. By clustering common dynamic behaviors derived from selectively blocked splicing reactions, SiMCAn was able to identify the signature conformations and dynamic behaviors of multiple ATP-dependent intermediates. In addition, it identified an open conformation adopted late in splicing by a 3′ splice-site mutant, invoking a mechanism for substrate proofreading. SiMCAn enables rapid interpretation of complex single-molecule behaviors and should prove useful for the comprehensive analysis of a plethora of dynamic cellular machines.

Conformational dynamics have a key role in every aspect of RNA biology, including RNA transcription, splicing and translation[1–3]. 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[2–4]. Single-molecule fluorescence approaches have recently emerged as a powerful toolset for dissecting the structural dynamics that form the foundation of biomolecular machines functioning at the nanometer scale[5–9]. For example, smFRET has been used to dissect spliceosome dynamics[5,6,10]. The spliceosome is a multi-megadalton ribonucleoprotein complex essential for the faithful removal of introns from eukaryotic precursor mRNAs (pre-mRNAs) during the two chemical steps of splicing (**Fig. 1a**)[11]. The architectural reorganization of the pre-mRNA substrate required to accommodate these two catalytic steps in a single active site is thought to be accompanied by substantial rearrangements that ensure substrate proofreading[12–15]. To explore these rearrangements, in previous studies[6,16] we labeled the efficiently splicing yeast pre-mRNA Ubc4 with the FRET pair Cy5 and Cy3 seven nucleotides upstream of the 5′ splice site (5′SS) and six nucleotides downstream of the branch point (BP), respectively. This yielded a substrate capable of detecting changes in intron conformation as a result of 5′SS and BP (un)docking (**Fig. 1a**,**b**) that we 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[5].

The quantitative methods available for an in-depth dissection of the dynamics observed in smFRET studies are still limited, however. In particular, the multistate, mostly asynchronous and often heterogeneous kinetics of many molecular machines, such as the spliceosome, mean that even with current state-of-the-art analysis, individual state transitions are rendered as independent stochastic events insufficient for an in-depth understanding of the underlying biological function. To extract additional information, several recent studies analyzed common smFRET metrics more thoroughly, specifically, FRET probability histograms and state-to-state transition kinetics[7]. For example, it has been demonstrated that in certain favorable cases interstate dynamics can be extracted from histograms through an analysis of photon arrival times and lifetimes[17]. In addition, state-to-state transition kinetics have been extracted through the use of clustering algorithms to identify distinct kinetic behaviors[18,19]. All of these approaches have focused on small data sets with two or three FRET states and limited dynamics; to date they have not been applied to more complex systems with higher numbers of states and complex kinetic networks examined under non-equilibrium conditions.

We present here a method, Single Molecule Cluster Analysis (SiMCAn), that utilizes hierarchical clustering as a means to group, sort and identify commonalities of smFRET trajectories fit using hidden Markov modeling (**Fig. 1c**,**d**). We used SiMCAn to characterize the pre-mRNA dynamics associated with the assembly and catalytic steps of the yeast spliceosome. SiMCAn reduces every single-molecule trajectory, regardless of its number of states, to an easily comparable unit of information that we refer to as the FRET similarity matrix (FSM). By leveraging hierarchical clustering techniques, we identified common dynamic behaviors

[1]Department of Chemistry, Single Molecule Analysis Group, University of Michigan, Ann Arbor, Michigan, USA. [2]Cellular and Molecular Biology, University of Michigan, Ann Arbor, Michigan, USA. [3]Biology Department, University of North Carolina, Chapel Hill, North Carolina, USA. [4]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California, USA. [5]Present addresses: Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA (M.R.B.); National Evolutionary Synthesis Center, Durham, North Carolina, USA (J.S.M.). [6]These authors contributed equally to this work. Correspondence should be addressed to N.G.W. (nwalter@umich.edu).

RECEIVED 4 SEPTEMBER 2014; ACCEPTED 10 AUGUST 2015; PUBLISHED ONLINE 28 SEPTEMBER 2015; DOI:10.1038/NMETH.3602

across 10,680 different Ubc4 pre-mRNA molecules. We accomplished 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 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 (3′SS) mutant. Our results establish SiMCAn as an effective approach for characterizing complex smFRET behaviors of dynamic cellular machines.

RESULTS

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 used tools for identifying state-to-state transitions in smFRET trajectories (**Fig. 1c**,**d**). HMM fits create challenges, however, in comparisons of trajectories with different states and kinetic properties across a variety of experimental conditions (**Supplementary Note 1**). Fitting all data with a single HMM, so that consistent state values are used across all trajectories, is one way to address these challenges[6,7]. Such an approach effectively imposes a single, preordained kinetic model on all molecules and experimental conditions, which might not be appropriate for highly complex systems such as the spliceosome.

SiMCAn presents a solution for sorting and identifying commonalities among large numbers of HMM-fitted smFRET trajectories by first binning each FRET state into one of ten evenly spaced FRET values (0.05–0.95, with increments of 0.10) (**Fig. 2a**). This binning enables the global analysis of a large data set with FRET values that 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 that describe the FRET states as well as the kinetics of transition between them (**Fig. 2a**). Each TP matrix is then combined with the occupancies of the individual FRET states to create an FSM (**Fig. 2a**). The Euclidean (ordinary) distance between FSMs provides a suitable weighted, information-rich metric by which to compare thousands of HMM-fitted smFRET trajectories using hierarchical clustering analysis (**Supplementary Note 1**), an agglomerative clustering technique that aims to group data of similar characteristics without the need for a preconceived experimental model or hypothesis[20,21] (Online Methods). The result of this clustering is a hierarchical tree, where each leaf on the tree represents the dynamics of an individual molecule and BPs indicate a split in the dynamic behavior of the group of molecules at a given level of coarseness (**Fig. 2b**). The number of clusters is determined using an iterative measurement of the intercluster distances and a modified *k*-means algorithm[22]. Each cluster will be represented using the average TP matrix, a random collection of traces and the probability distribution of FRET states in the cluster (**Fig. 2c**).

Validation of SiMCAn using simulated data sets

To evaluate whether SiMCAn is able to correctly identify and segregate HMM-fitted trajectories with known FRET states, we applied it first to a simulated data set 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 (**Supplementary Fig. 1a**), with average rate constants of $k_{0.15\rightarrow0.45}$ = 0.54 s^{-1}, $k_{0.45\rightarrow0.15}$ = 0.54 s^{-1}, $k_{0.15\rightarrow0.85}$ = 0.54 s^{-1} and $k_{0.85\rightarrow0.15}$ = 0.54 s^{-1}. Using the intercluster distances and modified *k*-means algorithm, SiMCAn

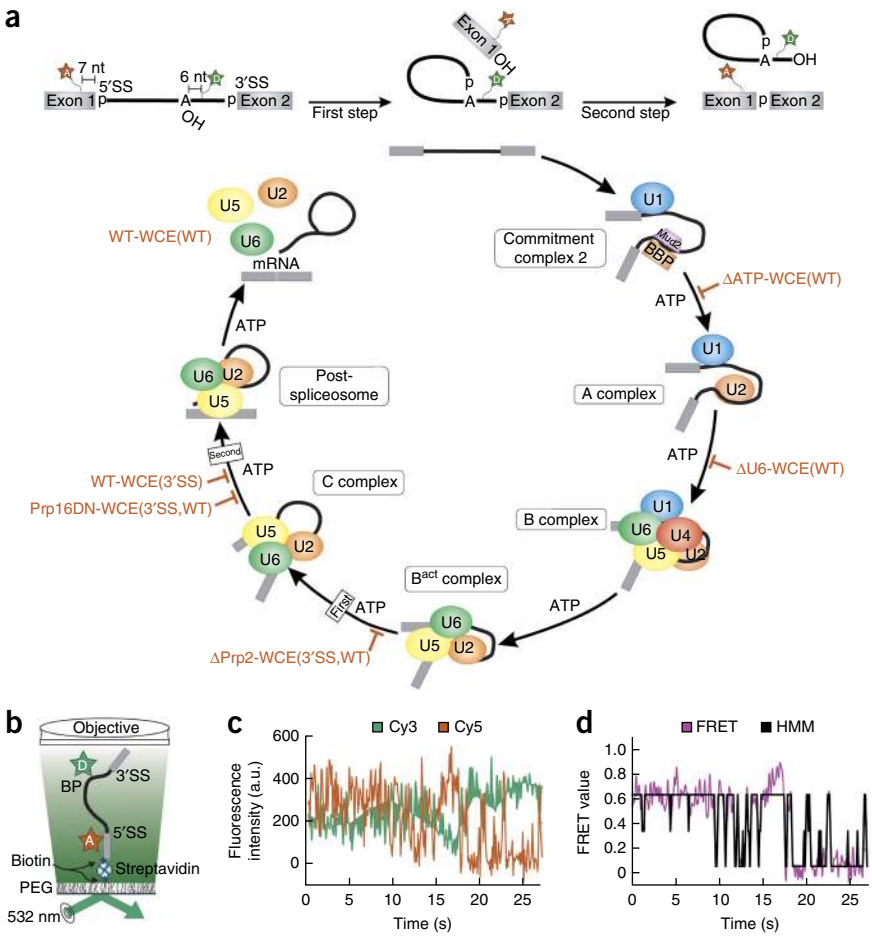

Figure 1 | smFRET analysis of pre-mRNA splicing using the HMM. (**a**) The fluorescent substrate used to monitor pre-mRNA dynamics contains Cy5 and Cy3 fluorophores seven nucleotides upstream of the 5′SS and six nucleotides downstream of the BP, respectively. Spliceosome assembly and catalysis are thought to progress in a stepwise manner, with ATP required at several steps of assembly. The biochemical and genetic stalls used in this study are indicated by orange blocks. (**b**) Prism-based total internal reflection fluorescence microscopy setup for smFRET. (**c**) Raw single-molecule time trace showing the anticorrelated donor (green) and acceptor (orange) intensities. (**d**) The corresponding FRET trace (purple) and the HMM trace as assigned by vbFRET (black).

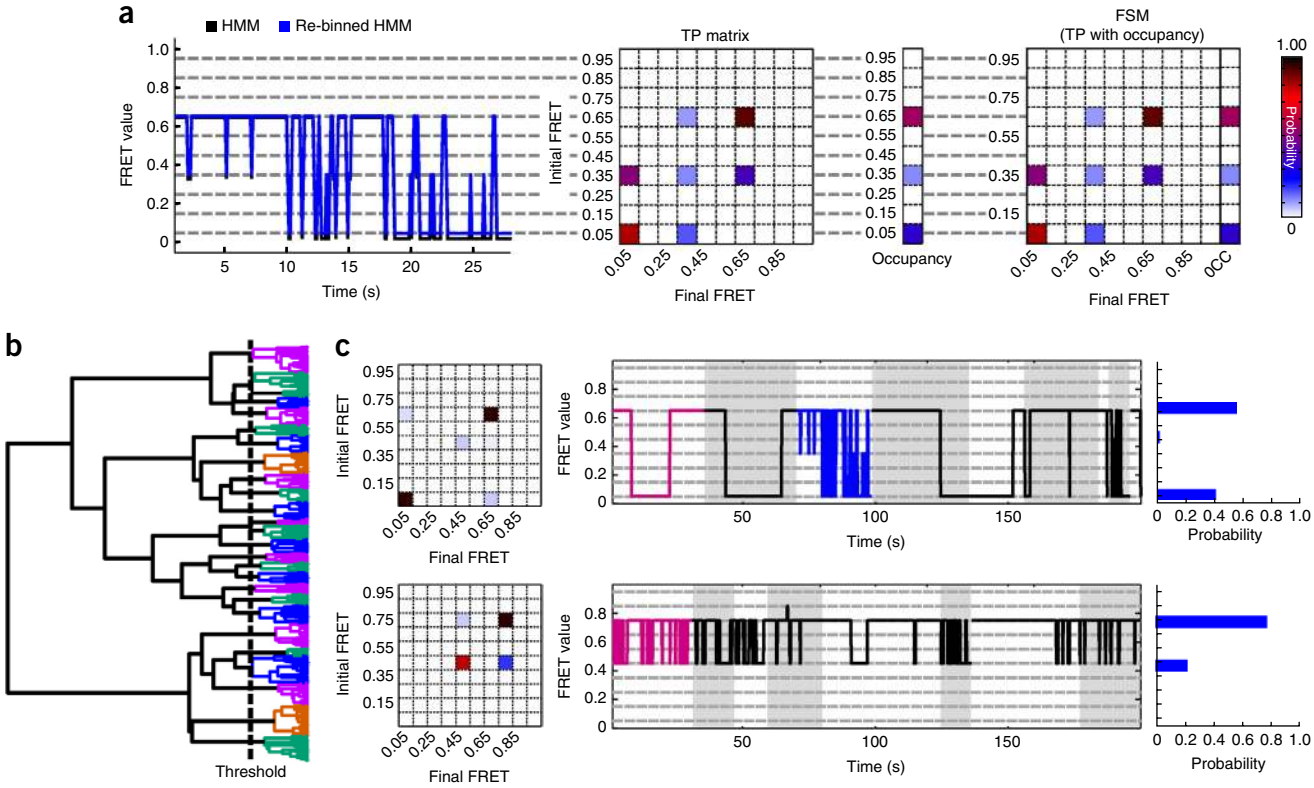

Figure 2 | SiMCAn workflow for sorting and clustering single-molecule-derived HMMs for common dynamic behaviors. (**a**) Left, assigned FRET trace before (black) and after (blue) reassignment to the closest of ten evenly spaced states (0.05–0.95 in increments of 0.10; gray dashed lines). Center, TP matrix corresponding to the re-binned FRET trace in the left-hand panel and occupancy values for each of the ten FRET values for the molecule traced in the left-hand panel. Right, FSM containing the TP matrix and FRET occupancies that describe the FRET states and transition kinetics between them for the molecule in **a**. (**b**) Hierarchical tree resulting from hierarchical clustering analysis using all 6,079 dynamic molecules. 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. Static molecules were identified and analyzed by SiMCAn separately. (**c**) Cluster description for 2 of the 25 dynamic clusters of the full splicing data set. Each representation shows the TP matrix of the cluster, the trace closest to the cluster center (magenta), up to 200 s of random (black) traces from the cluster and the probability of FRET states within the cluster. The highlighted blue trace in the top right panel represents the example trace used in **a**. Gray and white shading in backgrounds demarcates individual trajectories in **c**.

properly identified and separated these two molecular behaviors (**Supplementary Fig. 1b**), demonstrating that FSMs can be clustered and distinguished on the basis of the identity of their FRET states. A second and more important feature of SiMCAn is the ability to segregate HMMs on the basis of differing kinetics. We analyzed a second set of 3,000 simulated HMMs possessing two FRET states of 0.15 and 0.75, with half designed to have identical interconversion rate constants of 0.54 s^{-1} and the other half transitioning much more slowly, with rate constants of 0.15 s^{-1} (**Supplementary Fig. 1c**). SiMCAn identified two clusters with distinct transition rate constants between the two states (**Supplementary Fig. 1d**). These results demonstrate SiMCAn's ability to differentiate HMM-fitted FRET trajectories on the basis of their FRET states and kinetics.

Validation of SiMCAn using purified spliceosomal complexes

To benchmark SiMCAn against a more complex experimental data set featuring multiple FRET states, numerous transition rate constants and the inherent experimental limitations (for example, signal noise and premature photobleaching), we chose to analyze a previously published data set collected during the Prp2-mediated conformational transition immediately before the first step of splicing[5]. Briefly, immobilized B^{act} complex containing FRET-labeled Ubc4 was monitored as it progressed through the B* to the C complex after the addition of recombinant proteins Prp2, Spp2 and Cwc25 (**Fig. 3a**). Only after exhaustive manual sorting were we able to identify distinct FRET state and kinetic signatures for the intermediate B^{act}, B* and C complexes (**Supplementary Fig. 2a**). In contrast, SiMCAn was able to rapidly (within minutes) and correctly identify these previously only manually identified[5] (**Supplementary Fig. 2b**) subpopulations of pre-mRNA molecules as follows.

The HMM-fitted FRET traces under B^{act}, B* and C complex conditions were combined and analyzed using SiMCAn to determine whether the analysis could recapitulate the manual annotation of these traces. Maximizing the intercluster distances while minimizing the intracluster distances using SiMCAn yielded nine dynamic and four static clusters that best fit the data (**Fig. 3b** and **Supplementary Fig. 3**). The data for these clusters were combined into a single bar graph to depict the fraction of molecules that occupied each cluster, which allowed for the identification of the most populated clusters under each experimental condition (**Fig. 3c**). In results similar to those of our previous analysis, a cluster of molecules adopting a static low-FRET state (0.3-S) was identified as dominant under B^{act} complex conditions (**Fig. 3c**), whereas a static high-FRET cluster (0.7-S) was most abundant

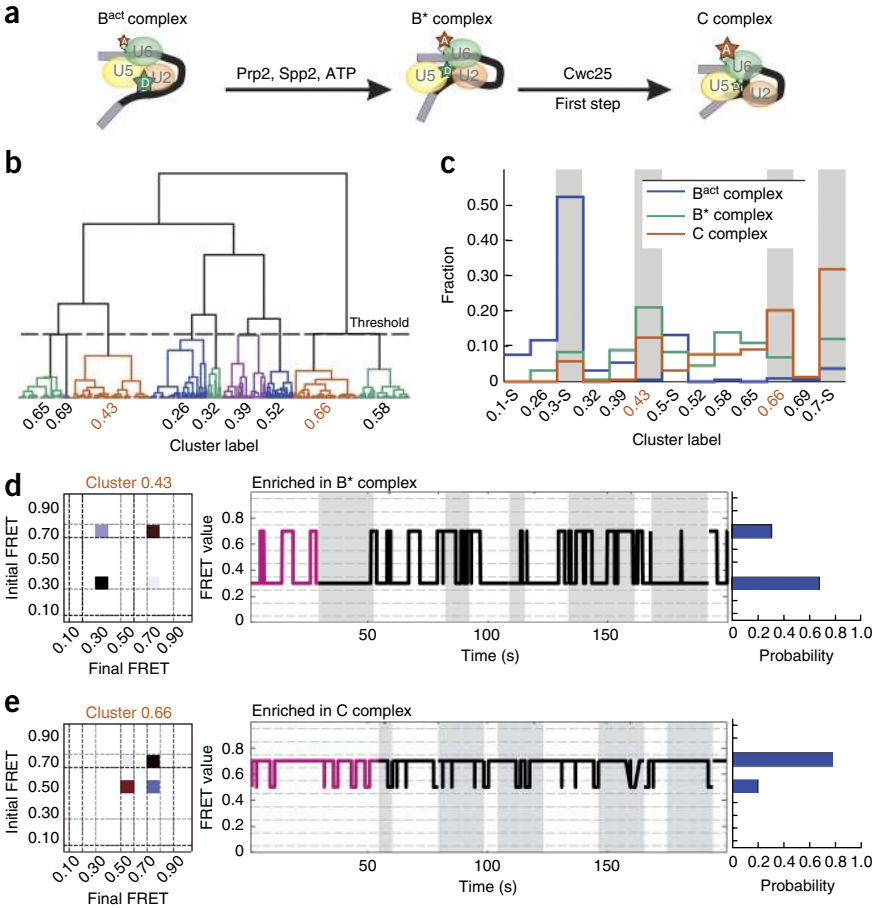

Figure 3 | Validation of SiMCAn using a previously analyzed data set describing the transition from the purified B^{act} complex to the C complex[5]. (**a**) Protein requirements for the transition from the B^{act} complex through B* to the C complex. (**b**) Hierarchical tree based on hierarchical clustering analysis of the dynamic molecules re-fit with FRET states of 0.1, 0.3, 0.5 and 0.7 (ref. 5). Static molecules were identified and analyzed by SiMCAn separately. (**c**) Cluster occupancy showing the fraction of molecules from each experimental condition that occupied the nine dynamic and four static clusters found using SiMCAn. Dynamic clusters are labeled by the weighted-average FRET value of the molecules in the cluster (for example, 0.2563), and static clusters are labeled according to the single state they describe (for example, 0.1-S). Gray bars highlight the most populated clusters occupied by each of the complexes. (**d**,**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 complexes 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).

under C complex conditions (**Fig. 3c**). In addition, SiMCAn identified two dynamic clusters that were increasingly populated under B* (cluster 0.43) and C (cluster 0.66) complex conditions (**Fig. 3c**). Cluster 0.43 contained molecules with a short-lived high-FRET state and longer dwell times in the low-FRET state that were most abundant under B* conditions (**Fig. 3d**). By contrast, cluster 0.66 contained molecules with a longer-lived high-FRET state featuring rapid excursions back to a mid-FRET state that were enriched upon the addition of Cwc25 to form the C complex (**Fig. 3e**), matching our previous manual analysis[5]. These results demonstrate that, when applied to a complex experimental data set, SiMCAn is able to segregate data efficiently on the basis of FRET states and differences in state-to-state interconversion kinetics to derive a biologically meaningful result.

Stalling of the spliceosome leads to distinct behaviors

Having established that SiMCAn identifies known dynamic behaviors in simulated (**Supplementary Fig. 1**) and experimental HMM-fitted smFRET trajectories (**Fig. 3**), we next used it on a new data set enriched for specific stages of splicing through the use of biochemical and genetic stalls for which no behaviors were known. We collected smFRET data after incubating FRET-labeled wild-type Ubc4 pre-mRNA with wild-type yeast whole-cell extract (WCE), allowing for spliceosomal assembly on and splicing of the fluorescent substrate (condition WT-WCE(WT); **Fig. 1a**). Time-course experiments were performed during which smFRET was recorded in time windows 0–8 min (early), 18–23 min (middle) and 33–40 min (late) after the addition of WCE. To assign dynamics to particular splicing intermediates without a need for cumbersome biochemical isolation, we chose to utilize eight mutations, and combinations thereof, known to allow for efficient accumulation of specific splicing intermediates in WCE (**Fig. 1a** and **Supplementary Table 1**). Blockage and release by reconstitution were verified by bulk *in vitro* splicing assays in yeast WCE (**Supplementary Fig. 4**). smFRET data for each stall were then acquired using the same time-lapse approach used for the WT-WCE(WT) condition. FRET probability distributions and transition occupancy density plots (**Supplementary Figs. 5** and **6**) were used to broadly summarize the behavior of hundreds of molecule trajectories per condition[7], confirming that the blocks led to different ensemble and time-averaged behaviors. However, this far more complex data set is not amenable to standard analysis techniques, as it includes a large number of traces, FRET states and transition-rate constants from splicing complexes stalled by mutation throughout the splicing cycle. It thus represents an ideal application for SiMCAn.

Identifying biologically defined dynamics using SiMCAn

Application of SiMCAn to this new data set allowed us to identify and cluster sets of molecules that shared common dynamic behaviors. Each of the 10,680 smFRET trajectories was first fit with an HMM using vbFRET[23], although any HMM fitting tool that satisfies the user's fitting preferences can be used. Prior to clustering, 4,601 static molecules were identified and analyzed separately. Hierarchical clustering of the remaining 6,079 dynamic molecules produced a tree that was pruned to a height of 25 distinct clusters (**Fig. 2b** and **Supplementary Fig. 7**), so that each cluster represented a unique dynamic behavior (**Fig. 2c** and **Supplementary Fig. 8**). Static clusters were

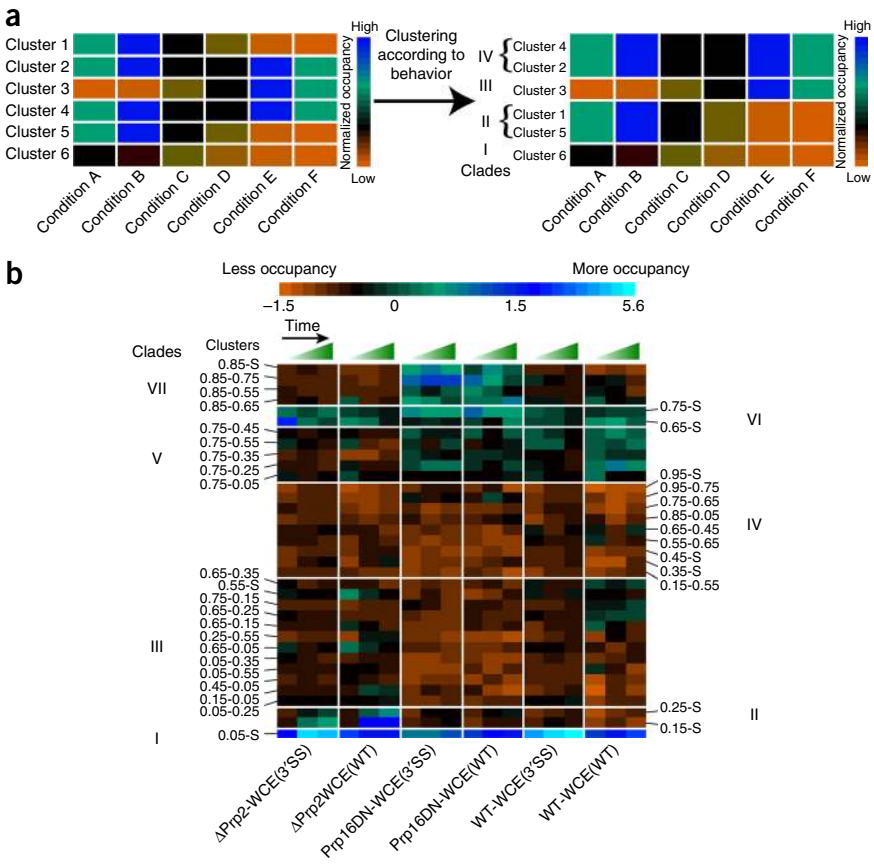

Figure 4 | Clustering of clusters to identify 'clades' of similar behavior. (**a**) Illustration of the second round of clustering to group the clusters by common occupancy patterns. In this example, six clusters (1–6) have been populated by six conditions (A–F). Each cluster has an occupancy pattern across the conditions as represented by a heat map, with high occupancy denoted by blue and low occupancy denoted by orange. After a second round of SiMCAn clustering, clusters with similar occupancies across the six conditions were grouped into clades (I–IV). (**b**) The second round of clustering with the 35 clusters from our experimental splicing data set revealed seven clades (I–VII) of clusters enriched in particular splicing complexes. The fraction of molecules in each cluster for each experimental condition at each time was normalized to a mean of zero with unit variance. Green and blue shading indicate increased occupancy of a particular cluster; orange 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.

named according to their sole FRET state (for example, 0.05-S), whereas dynamic cluster names were assigned on the basis of the first and second most occupied FRET states in the cluster (for example, cluster 0.65-0.05 primarily occupied 0.65 and 0.05 FRET states). Bootstrap analysis based on the 25 SiMCAn-identified clusters confirmed the ability to identify input HMMs from increasingly complex data sets and showed that the SiMCAn-identified clusters for the large experimental data set captured the molecular behaviors exhaustively (**Supplementary Fig. 9**).

We next sought to identify clusters whose occupancies were similarly enriched or depleted for the same group of conditions, that is, clusters that followed a similar pattern of high and low occupancies across conditions, such that they could be grouped into a 'clade' through a second round of hierarchical clustering (**Fig. 4a**). After applying this second level of SiMCAn to the full data set, we obtained a tree height of seven clades (**Supplementary Fig. 10**) that allowed for the identification of clusters representative of particular splicing conditions, most naturally capturing the changes in dynamic behavior expected to occur as pre-mRNA progresses through the splicing cycle (**Fig. 4b** and **Supplementary Fig. 11**). A bar graph of all 35 (25 dynamic and

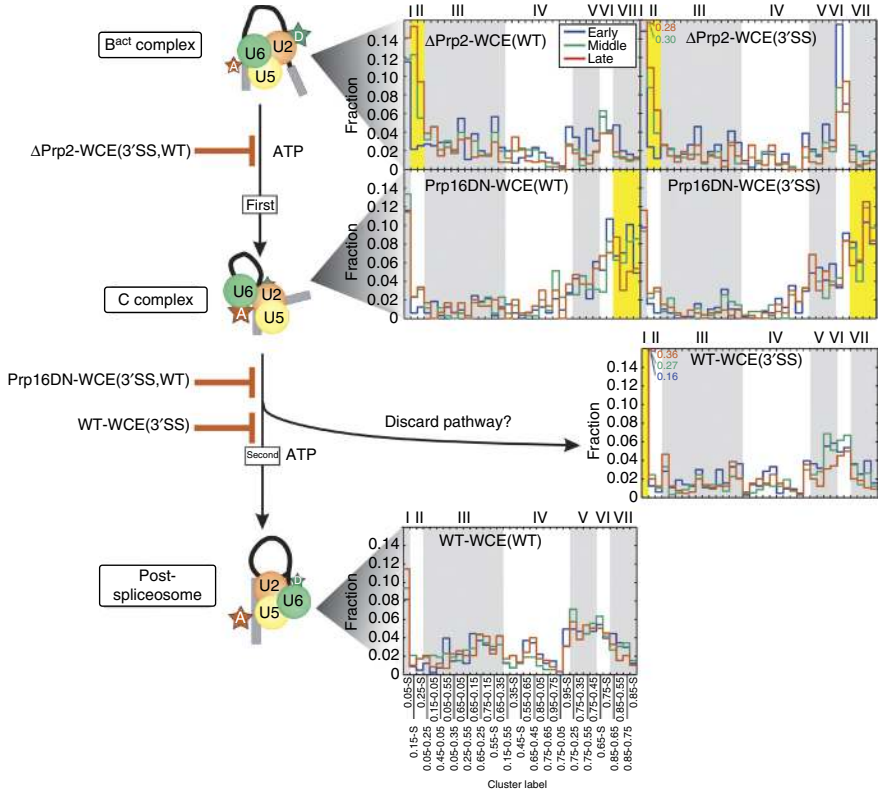

Figure 5 | Cluster-occupancy histogram showing the raw fraction of molecules occupying each cluster for the late assembly stages of the splicing cycle. Gray and white shading in backgrounds demarcates the clusters (bottom) composing each of the seven clades (top). Clusters with occupancy characteristic of a specified condition are highlighted in yellow.

10 static) clusters showed the extent to which each cluster contributed to the overall dynamics for each condition (**Fig. 5** and **Supplementary Figs. 12** and **13**). Statistical analysis showed that the average length of molecules in each cluster was similar, indicating that SiMCAn does not segregate by trace length (**Supplementary Fig. 14** and **Supplementary Table 2**).

Characterization of pre– and post–first step blocks

SiMCAn revealed a disperse set of dynamics and cluster occupancies in the early splicing conditions ΔATP-WCE(WT) and ΔU6-WCE(WT) that stalled at commitment complex 2 and the A complex, respectively (**Supplementary Figs. 13** and **15**). It also identified a time-dependent increase in clade I upon A-complex formation (**Supplementary Note 2**). This low-FRET behavior has been proposed to be sustained upon incorporation of the U5·U4/U6 tri-snRNP (small nuclear ribonucleoprotein) during B complex formation[10] (**Fig. 1**). In our corresponding data sets for conditions ΔPrp2-WCE(WT) and ΔPrp2-WCE(3SS), known[24,25] to enrich the activated spliceosome B^{act}, SiMCAn recognized a pair of static clusters, 0.25-S and 0.15-S, that were overrepresented and thus grouped to form clade II (**Figs. 4** and **5**). These clusters represented molecules that were stalled in a static low-FRET B^{act} conformation before activation of Prp2's ATPase activity and were similar to those previously determined[5] using an isolated B^{act} complex lacking free extract (**Fig. 3**). 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 was not resolvable in the FRET histograms (**Supplementary Fig. 5**). In addition to the static clusters of clade II, the dynamic cluster 0.05-0.25 (**Supplementary Fig. 16a**) was 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 the findings under Prp2 depletion, SiMCAn identified clade VII as particularly enriched upon the addition of recombinant Prp16 dominant-negative mutant ATPase (conditions Prp16DN-WCE(WT) and Prp16DN-WCE(WT)), known to stall splicing in the post–first step C complex[5,26,27] (**Figs. 4** and **5** and **Supplementary Figs. 16b** and **17**). In this clade were static cluster 0.85-S and three dynamic clusters, all containing the 0.85 FRET state (**Fig. 6**), which is distinct from the 0.75-S/0.65-S conformational state of clade VI enriched in early splicing intermediates (**Supplementary Fig. 13**). The dynamics of the clusters enriched at the Prp16DN stage indicated a preference for the 0.85 high-FRET state (**Fig. 6b**), which suggested that we were 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[5]. Although the ΔPrp2-WCE(3′SS) stall did show a delay in B^{act} complex

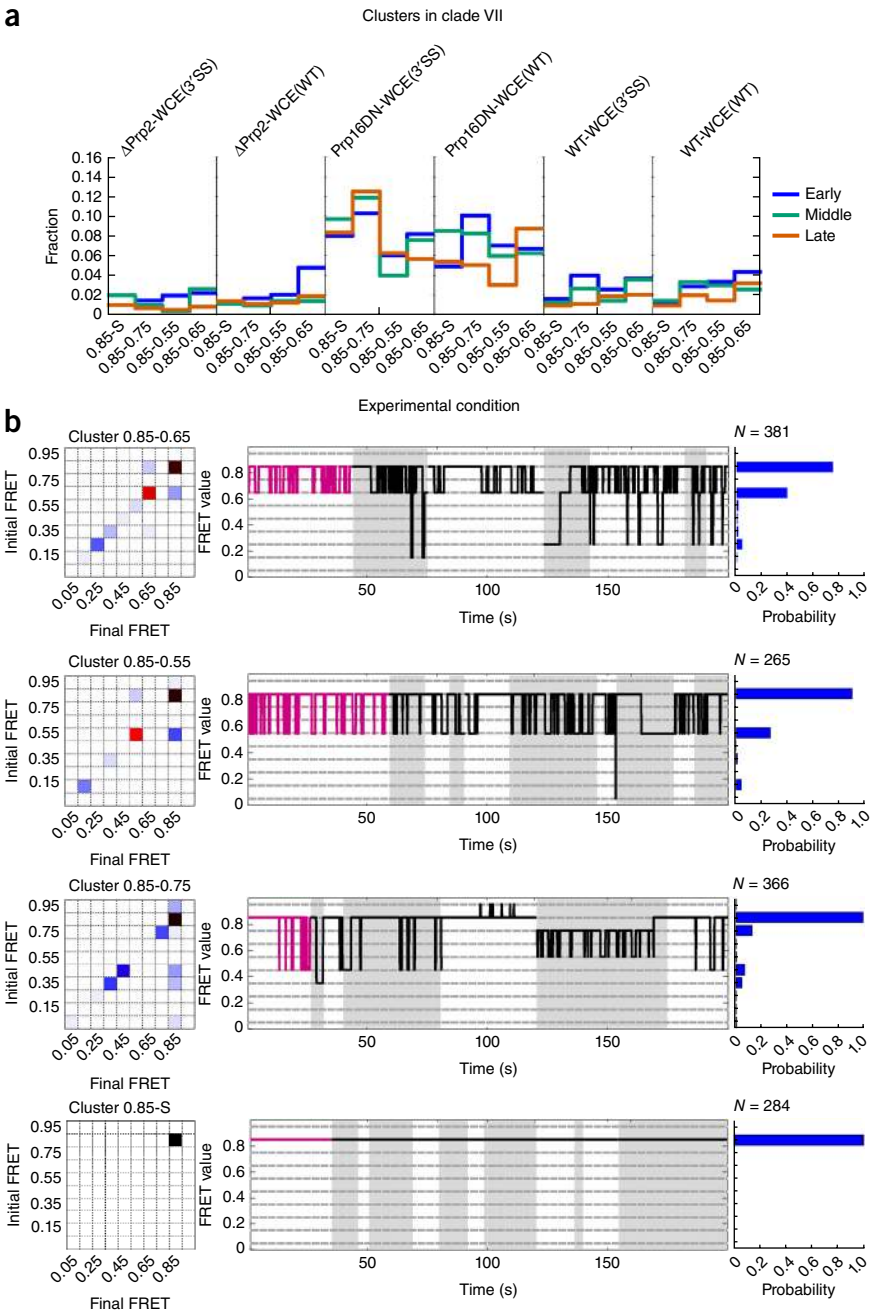

Figure 6 | Dynamic clusters of clade VII enriched in the Prp16DN-WCE conditions showed repeated excursions from the 0.85 state to lower FRET states. (**a**) Fraction of molecules in each late assembly stage for the clusters of clade VII. (**b**) Cluster description for each of the four clusters in clade 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). Gray and white shading in backgrounds demarcates individual trajectories.

formation (**Supplementary Note 3**), 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[28].

A 3′SS mutant undocks late in spliceosome assembly

Finally, SiMCAn identified differences in smFRET behavior between the wild-type and 3′SS mutant substrates after incubation with wild-type WCE containing no blocks (WT-WCE(WT) and WT-WCE(3′SS), respectively), thus allowing for unabated assembly toward 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 (**Supplementary Fig. 4**). As both substrates progressed through most of the splicing cycle, it is not surprising that SiMCAn revealed a similar set of sampled pre-mRNA conformations (**Fig. 5**). However, over time the 3′SS adopted an increasingly dominant 0.05-S cluster (**Fig. 5** and **Supplementary Fig. 18**), indicating a large separation of the 5′SS and BP not found in the Prp16DN-WCE(3′SS) data set. This 0.05-S state was thus stabilized to a much greater extent in the 3′SS mutant than in the wild-type 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 was detected. Our data suggest that the 3′SS either is 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[29,30]. 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 4**)[13,31]. Taken as a whole, our SiMCAn analysis is consistent with the hypothesis that the lack of a proper 3′SS sequence marker may trigger proofreading against a substrate that is not kinetically competent for the second step of splicing by undocking from the active site.

DISCUSSION

We show that SiMCAn reveals unique dynamic properties associated with specific splicing-cycle intermediates that cannot be identified using classical smFRET analysis (**Supplementary Figs. 5** and **6**). Because SiMCAn does not make assumptions about the heterogeneity or completeness of the underlying biochemical reactions, it allows one to identify consistent molecular behaviors in a model-free fashion (**Supplementary Note 1**). 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 of splicing when the spliceosome encounters a 3′SS mutation (**Fig. 5**). SiMCAn thus can use exploratory data sets 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.

SiMCAn was born out of the necessity to classify common kinetic behaviors over a broad range of experimental states. The construction of hierarchical trees from disparate sets of data is the basis of most phylogenetic inference, and the methods presented here are inspired from evolutionary analysis[32]. The clades identified by SiMCAn allowed 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 investigating complex systems, such as the spliceosome, is that it does not allow for the unambiguous definition of conformations from FRET states. In a simpler system, such as the P4–P6 subdomain of the *Tetrahymena thermophila* group I intron, docking and/or undocking of the GNRA tetraloop can be assigned to specific FRET values, which enables the development of an unambiguous kinetic model[19]. If combined with SiMCAn, emerging approaches involving multiple probes, such as coincidence analysis of colocalization single-molecule spectroscopy[33], could resolve this ambiguity and facilitate the development of a complete kinetic model of the eukaryotic splicing cycle. Furthermore, as point detector–mediated photon counting becomes more high-throughput, these methods should introduce a substantial improvement in time resolution and allow a detailed description of shot-noise-limited FRET efficiency distributions[17].

In summary, our results demonstrate that SiMCAn is a powerful tool for the unbiased extraction of FRET states and kinetics from complex smFRET data sets. Beyond the identification of FRET states, SiMCAn helps distinguish molecules with similar FRET levels but differing rates of interconversion. With an additional layer of clustering based on the occupancy of behaviors across a systematic set of experimental conditions with known effects, the method enables the identification of common and distinct behaviors among large numbers of single molecules. Thus 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 data set, allowing for unprecedented in-depth analyses of the dynamics of complex biomolecular machines.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors 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) for providing the dominant-negative Prp16 protein expression plasmid; N.N. Vo for compiling all Matlab scripts of SiMCAn into a GUI; and C. Guthrie, D.R. Semlow, J.P. Staley and A.A. Hoskins for providing valuable comments on the manuscript. The authors acknowledge funding from the US National Institutes of Health (grant R01GM098023 to N.G.W. and J.A.), the National Heart, Lung and Blood Institute (grant R01HL111527-01 to A.L.) and the National Science Foundation through the National Evolutionary Synthesis Center (NESCent) (grant NSF\#EF-0905606 to J.S.M.).

AUTHOR CONTRIBUTIONS

M.L.K. and R.K. performed *in vitro* splicing verification assays. M.L.K. and M.R.B. performed single-molecule experiments and performed data analysis. M.L.K. expressed and purified the Prp16DN protein. M.R.B. and J.S.M. wrote and developed the Matlab scripts for SiMCAn. M.L.K. prepared all fluorescent substrates and yeast whole-cell extracts. M.R.B., J.S.M., M.L.K., J.A., A.L. and N.G.W. jointly wrote the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Pitchiaya, S., Heinicke, L.A., Custer, T.C. & Walter, N.G. Single molecule fluorescence approaches shed light on intracellular RNAs. *Chem. Rev.* **114**, 3224–3265 (2014).
2. Mustoe, A.M., Brooks, C.L. & Al-Hashimi, H.M. Hierarchy of RNA functional dynamics. *Annu. Rev. Biochem.* **83**, 441–466 (2014).
3. Al-Hashimi, H.M. & Walter, N.G. RNA dynamics: it is about time. *Curr. Opin. Struct. Biol.* **18**, 321–329 (2008).
4. Cruz, J.A. & Westhof, E. The dynamic landscapes of RNA architecture. *Cell* **136**, 604–609 (2009).
5. 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).
6. Abelson, J. *et al.* Conformational dynamics of single pre-mRNA molecules during *in vitro* splicing. *Nat. Struct. Mol. Biol.* **17**, 504–512 (2010).
7. Blanco, M. & Walter, N.G. Analysis of complex single-molecule FRET time trajectories. *Methods Enzymol.* **472**, 153–178 (2010).
8. 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).
9. Walter, N.G. & Bustamante, C. Introduction to single molecule imaging and mechanics: seeing and touching molecules one at a time. *Chem. Rev.* **114**, 3069–3071 (2014).
10. Crawford, D.J. *et al.* Single-molecule colocalization FRET evidence that spliceosome activation precedes stable approach of 5′ splice site and branch site. *Proc. Natl. Acad. Sci. USA* **110**, 6783–6788 (2013).
11. 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).
12. Egecioglu, D.E. & Chanfreau, G. Proofreading and spellchecking: a two-tier strategy for pre-mRNA splicing quality control. *RNA* **17**, 383–389 (2011).
13. Semlow, D.R. & Staley, J.P. Staying on message: ensuring fidelity in pre-mRNA splicing. *Trends Biochem. Sci.* **37**, 263–273 (2012).
14. Staley, J.P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315–326 (1998).
15. Wahl, M.C., Will, C.L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701–718 (2009).
16. Abelson, J., Hadjivassiliou, H. & Guthrie, C. Preparation of fluorescent pre-mRNA substrates for an smFRET study of pre-mRNA splicing in yeast. *Methods Enzymol.* **472**, 31–40 (2010).
17. Gopich, I.V. & Szabo, A. Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET. *Proc. Natl. Acad. Sci. USA* **109**, 7747–7752 (2012).
18. Keller, B.G. *et al.* Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models. *J. Am. Chem. Soc.* **136**, 4534–4543 (2014).
19. 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).
20. 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).
21. Mall, R., Langone, R. & Suykens, J.A. Multilevel hierarchical kernel spectral clustering for real-life large scale complex networks. *PLoS One* **9**, e99966 (2014).
22. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Series B Stat. Methodol.* **63**, 411–423 (2001).
23. Bronson, J.E. *et al.* 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).
24. 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).
25. Warkocki, Z. *et al.* Reconstitution of both steps of *Saccharomyces cerevisiae* splicing with purified spliceosomal components. *Nat. Struct. Mol. Biol.* **16**, 1237–1243 (2009).
26. 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).
27. 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).
28. Rymond, B.C. & Rosbash, M. Cleavage of 5′ splice site and lariat formation are independent of 3′ splice site in yeast mRNA splicing. *Nature* **317**, 735–737 (1985).
29. 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).
30. 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).
31. 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).
32. Woese, C.R. & Fox, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**, 5088–5090 (1977).
33. Hoskins, A.A. *et al.* Ordered and dynamic assembly of single spliceosomes. *Science* **331**, 1289–1295 (2011).

ONLINE METHODS

Synthesis of pre-mRNA substrates. The Ubc4 pre-mRNA substrates used in this study (**Supplementary Table 3**) were synthesized as previously described[6]. 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. In the 3′SS mutant, the guanines at positions 115 and 117 on the 3′ segment were replaced with cytosines. We coupled the 5′ and 3′ fragments to Cy5 and Cy3 N-hydroxysuccinimidyl ester (GE Healthcare), respectively, by resuspending 4 nmol of RNA in 40 µl of 0.1 M sodium bicarbonate buffer, pH 9.0, and incubating it for 30 min at 60 °C with the proper dye pack dissolved in dimethyl sulfoxide. The conjugated fragments were ethanol precipitated and washed with 70% (vol/vol) ethanol to remove unconjugated dye. Unlabeled RNA was removed by purification on benzoylated naphthoylated DEAE–cellulose (Sigma) that was washed with 1 M NaCl containing 5% (vol/vol) ethanol. Fully labeled RNA fragments were eluted with 1.5 M NaCl containing 20% (vol/vol) ethanol and further precipitated to remove excess salt. Labeled fragments were combined with an equal molar amount of DNA splint (**Supplementary Table 3**) and ligated by incubation with RNA Ligase 1 (New England BioLabs) for 4 h at 37 °C as described[6,16]. Full-length, labeled Ubc4 was then purified on a denaturing 7 M urea, 15% (wt/vol) polyacrylamide gel. Incubation of the fluorophore-labeled Ubc4 substrate with splicing buffer alone and no spliceosomal components revealed a dominant high-FRET peak with a smaller low-FRET population featuring very little zero FRET (**Supplementary Fig. 19**), as expected[6,7].

Preparation of yeast whole-cell extract. Splicing active 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[6,34]. Briefly, cells were grown in yeast extract–peptone–dextrose medium to an OD_{600} 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% (vol/vol) 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 r.p.m. in a type-45 Ti Beckman rotor. The supernatant was then centrifuged at 37,000 r.p.m. 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% (vol/vol) glycerol, 0.1 mM PMSF, and 0.25 mM benzamidine with one buffer exchange.

Accumulation of splicing complexes. Supplementary Table 1 describes all experimental conditions by identifying the substrate and WCE used along with the complex formed. We confirmed all splicing products 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 $K_i(PO_4)$, 20 mM KCl, 8% (vol/vol) glycerol, 3% (wt/vol) PEG, 0.5 mM DTT) and 40% (vol/vol) WCE at 25 °C for 40 min. Products were analyzed by separation on a 7 M urea, 15% (wt/vol) polyacrylamide gel and scanned on a Typhoon variable-mode imager (GE Healthcare; **Supplementary Fig. 4**). We performed ATP depletion by pre-incubating WCE with 1 mM glucose at 25 °C for 10 min before incubating it with splicing buffer and substrate. Endogenous U6 snRNA was depleted by pre-incubation of WCE with 300 nM D1 oligodeoxynucleotide (**Supplementary Table 3**) in splicing buffer, 50% (vol/vol) WCE, and 2 mM ATP at 33 °C for 30 min before incubation with substrate. We induced knockdown of endogenous Prp2 by heating *prp2-1 cef1-TAP* WCE to 37 °C for 40 min before incubating it with splicing buffer, ATP, and pre-mRNA substrate. Endogenous Prp16 was inactivated with 100 nmol of a Prp16 dominant-negative mutant (Prp16DN; K379A) added to the BJ2168 WCE for 10 min before incubation with splicing buffer, 2mM ATP, and pre-mRNA substrate. On-slide splicing assays were performed in the same way as the *in vitro* splicing assays with the exception that all materials were combined before reaction mixtures were flowed onto a substrate-coated, PEG-passivated slide using established procedures[5,6].

Single-molecule FRET. Single-molecule FRET was carried out in the same manner as previously described[5,6]. Using a prism-based total internal reflection fluorescence microscope[8,35,36], we collected data from single molecules incubated under the desired conditions (**Supplementary Table 1**). 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 the addition of WCE. The donor (Cy3) near the BP adenosine was excited with a 532-nm laser for 100 s, and then the Cy5 acceptor near the 5′SS was directly excited with a 635-nm laser for another 100 s. The resulting emission was recorded at 100-ms time resolution with a Princeton Instruments I-PentaMAX intensified CCD (charge-coupled device) camera. Molecules selected for further analysis by SiMCAn were required to last longer than 3 s before photobleaching of Cy3, show anti-correlated changes in Cy3 and Cy5 intensity, undergo single-step photobleaching, and still contain active Cy5 fluorophore at the time of their direct excitation. We calculated the FRET ratio by dividing the intensity of the acceptor emission by the total emission from both donor and acceptor. Each individual FRET trace was fitted with an individual HMM with up to ten states using vbFRET[23] in Mathwork's Matlab environment, with no assumptions about the values or distributions; in principle, any HMM-fitted trajectories could be used (generated by vbFRET, HaMMy, QuB, etc.)[7]. Regardless of the HMM software used, a certain degree of uncertainty in the number of FRET states and transitions among those states will be present in the data because of the noise associated with smFRET analysis. However, improvement of HMM analysis techniques is not the focus of this paper.

SiMCAn. The HMM-idealized data were assigned to the closest of ten evenly spaced FRET states (0.05–0.95, with an increment of 0.10 as our resolution limit). Traces of less than 3 s (30 frames) in length were discarded, and a TP matrix was constructed for each of the remaining molecule traces. Each TP matrix was then combined with the vector describing the percentage of the trace that occupied each FRET state to create a FRET similarity matrix (FSM) such that $FSM(i, j) = (TP(i - 1, j), P(n, j))$, where $i = 1 \ldots, n + 1$

and $j = 1 \ldots, n$. The FSMs were divided into categories containing static traces and dynamic traces, with the dynamic traces identified and characterized by having at least one FRET transition between two FRET states. Static traces were identified automatically on the basis of their unique signatures with just a single FRET value and were kept separate for the remaining analysis. Static molecules could arise as a result of fluorophores photobleaching before a transition took place. Alternatively, formation of a particular complex may lead to a very stable, unchanging conformation that results in a single (static) FRET state. The FSMs corresponding to dynamic traces were used as input for a hierarchical clustering analysis performed by Matlab (**Supplementary Software**) that calculates the distance between FSMs using the Euclidean (ordinary) distance. 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 and 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 intercluster distances and a modified *k*-means algorithm. The specific cutoff was chosen as the first point where randomly assigned traces had a higher intercluster distance than the hierarchical clustering, which provided the best option among several for determining an optimal cluster selection. 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 (**Supplementary Software**). For each experimental condition, we calculated the fraction of molecules in each SiMCAn-identified cluster by dividing the number of molecules in that condition assigned to each cluster by the total number of molecules in that condition. We used the occupancy in all the clusters as a new similarity matrix to compute the distance between each SiMCAn cluster using Euclidean-distance measurement. Clades were generated via the iterative *k*-means approach used in SiMCAn, with the aim of generating groups of clusters whose occupancy patterns across conditions were most alike (as measured by Euclidean distance). A detailed description of the mathematical algorithm is provided in **Supplementary Note 5**.

Generation of the simulated data sets. Artificial HMMs containing the distinctions of interest were used to generate traces of 10^6 time-step length for each of four clusters. These traces were used to generate 1,500 subtraces with the starting points uniformly selected along the full trace and the length determined by a Poisson distribution with a λ of 100. The resulting traces were treated exactly like experimentally acquired data fit by vbFRET for analysis by SiMCAn. The simulation data are available at https://app.box.com/s/v64wet7ixlkr7cb96ky9ij5pzqjlxnj3.

34. Stevens, S.W. & Abelson, J. Yeast pre-mRNA splicing: methods, mechanisms, and machinery. *Methods Enzymol.* **351**, 200–220 (2002).
35. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5**, 507–516 (2008).
36. 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.* **88**, 1965–1985 (2014).

doi:10.1038/nmeth.3602