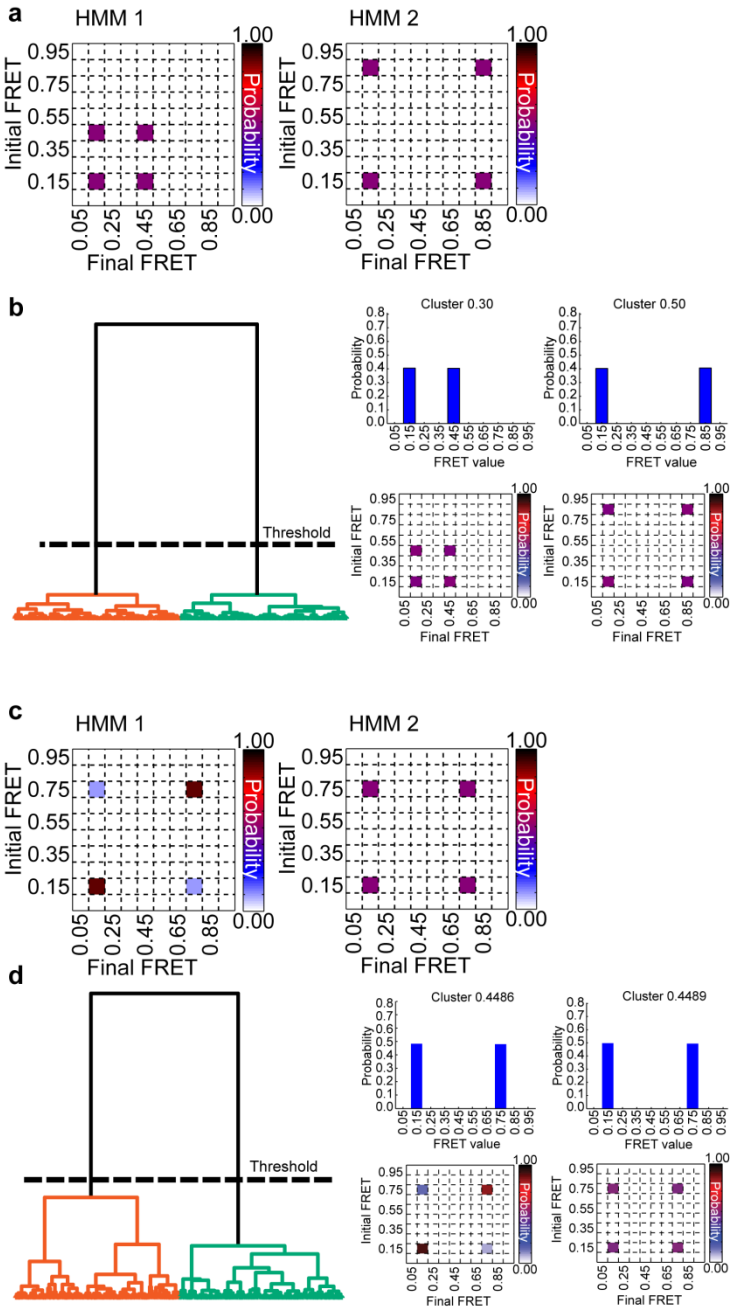
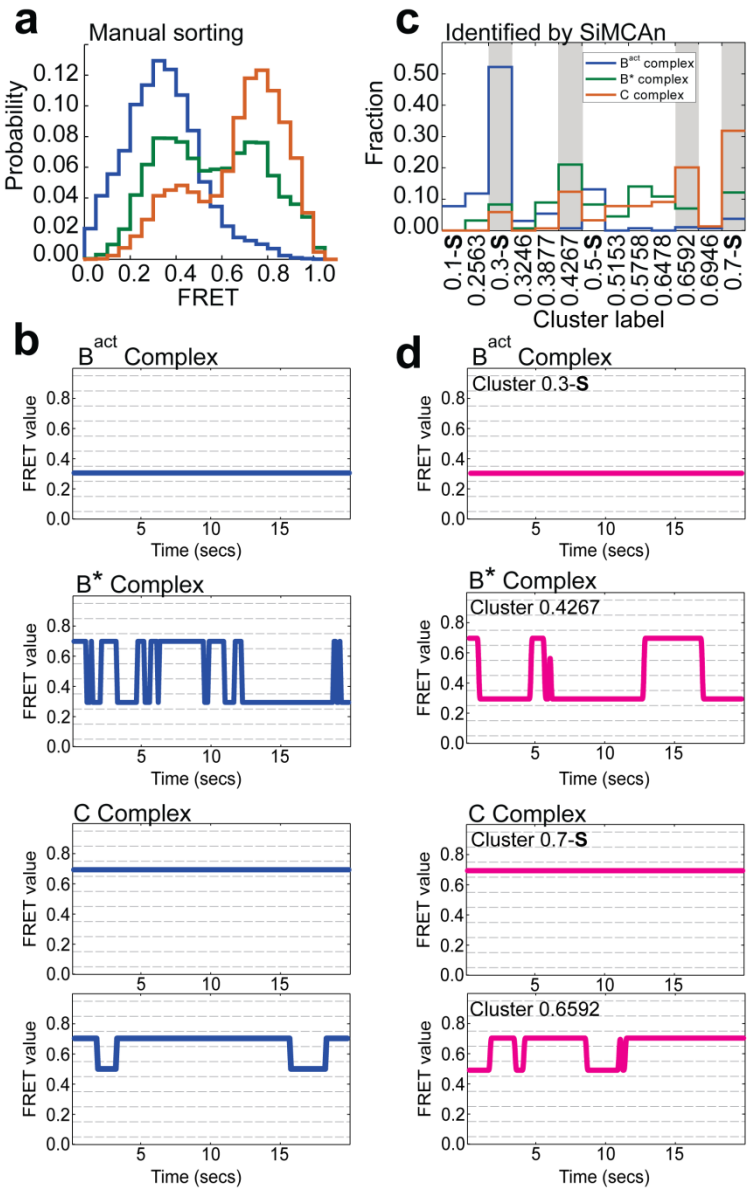
Blanco et al., Supplementary Figure 1



Supplementary Figure 1. Clustering of simulated datasets. (a) Transition probability (TP) matrices possessing one shared FRET state (0.15) and one differing FRET state (0.45 or 0.85) that were used to generate the 1500 random traces for clustering by SiMCAn. **(b)** Hierarchical tree showing the two cluster threshold found by SiMCAn and the two resulting cluster probability histograms and TP matrices. **(c)** TP matrices possessing the same two FRET states but different rates of interconversion used to generate the 1500 random traces for clustering by SiMCAn. **(d)** Hierarchical tree showing the three cluster threshold found by SiMCAn and the three resulting cluster probability histograms and TP matrices.

Nature Methods: doi:10.1038/nmeth.3602

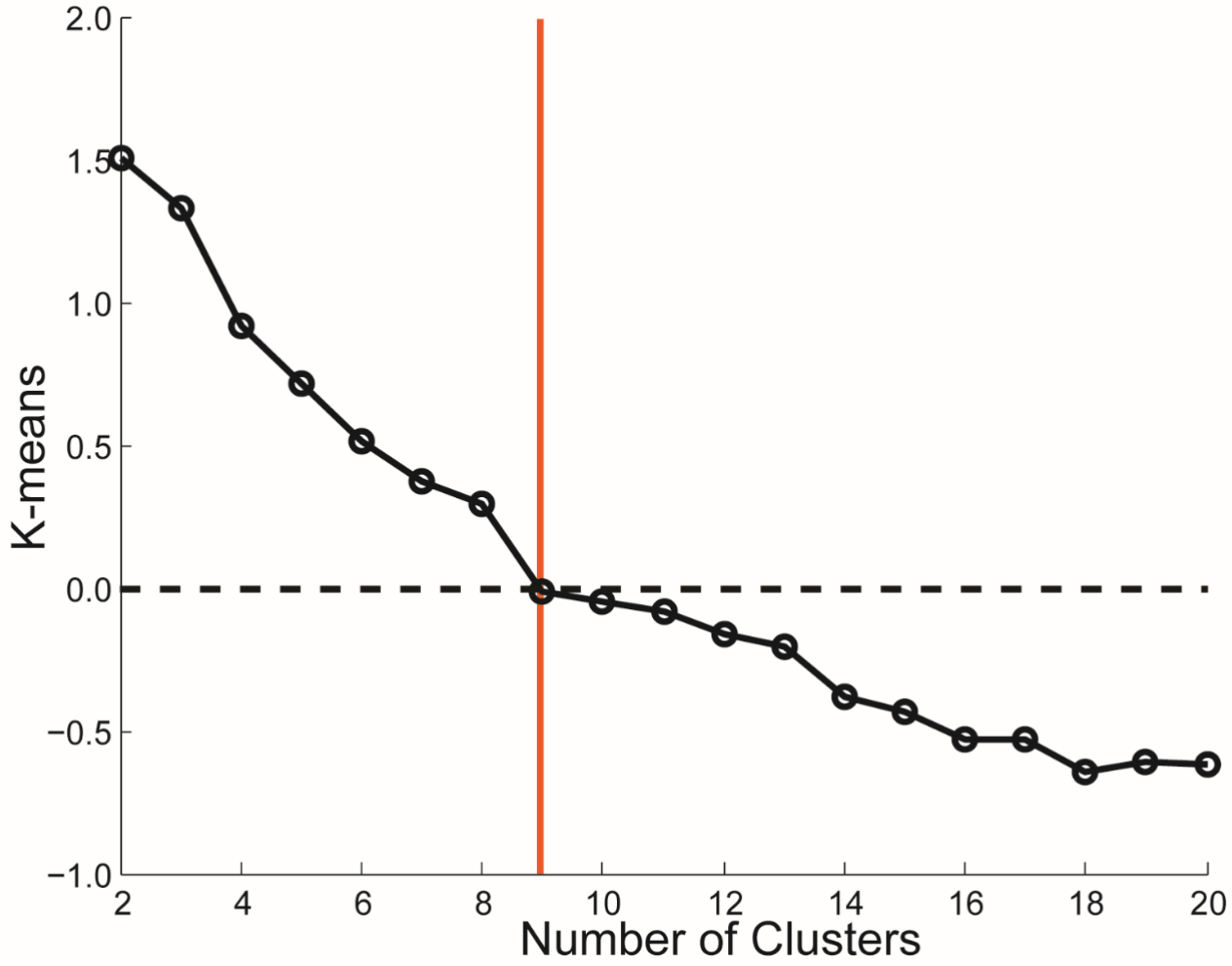
Blanco et al., Supplementary Figure 2



Supplementary Figure 2. Validation of SiMCAn using previously analyzed data describing the transition of the B^{act} complex through the C complex and first step of splicing. (a) FRET probability analysis for molecules in the B^{act} (blue), B* (green), and C complexes (red). **(b)** Visually most representative FRET trajectories describing the dominant behavior of molecules in each complex found through manual sorting of traces **(c)** Cluster occupancy bar graph showing the fraction of molecules from each experimental condition that occupy the 9 dynamic and 4 static clusters found using SiMCAn. Dynamic clusters were labeled by the weighted average FRET value of the molecules within the cluster (e.g., 0.2563) while static clusters are labeled by the single state they describe (e.g., 0.1-**S**). Grey bars highlight the most populated clusters occupied by each of the complexes. **(d)** The smFRET trajectories found using SiMCAn that are most similar to the cluster center of the four indicated clusters that describe each of the splicing complexes.

Nature Methods: doi:10.1038/nmeth.3602

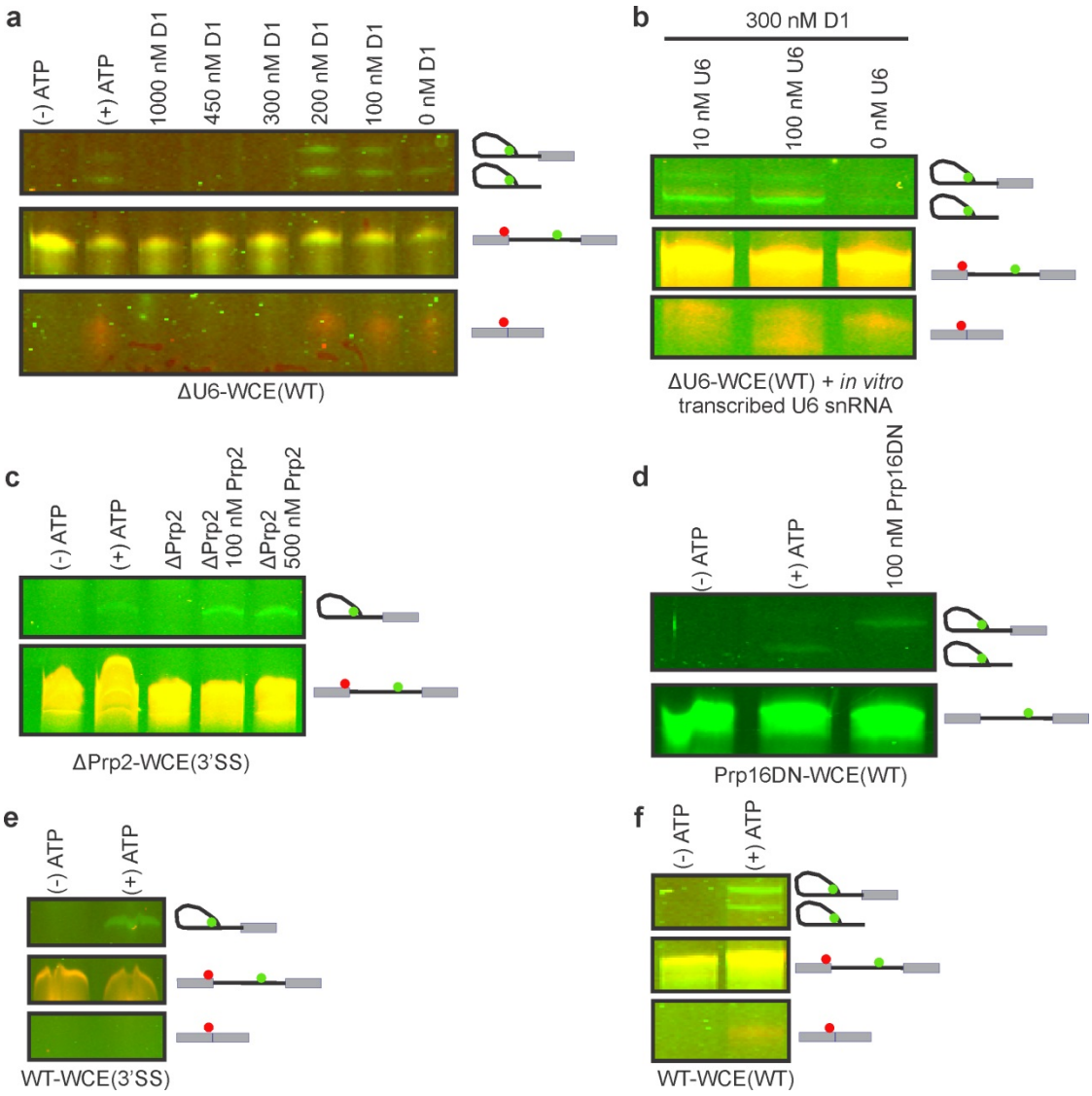
Blanco et al., Supplementary Figure 3



Supplementary Figure 3. Iterative measurement of inter-cluster distances using a modified k-means algorithm utilized to determine the number of clusters that best describes the previously analyzed B^{act} dataset[1].

Nature Methods: doi:10.1038/nmeth.3602

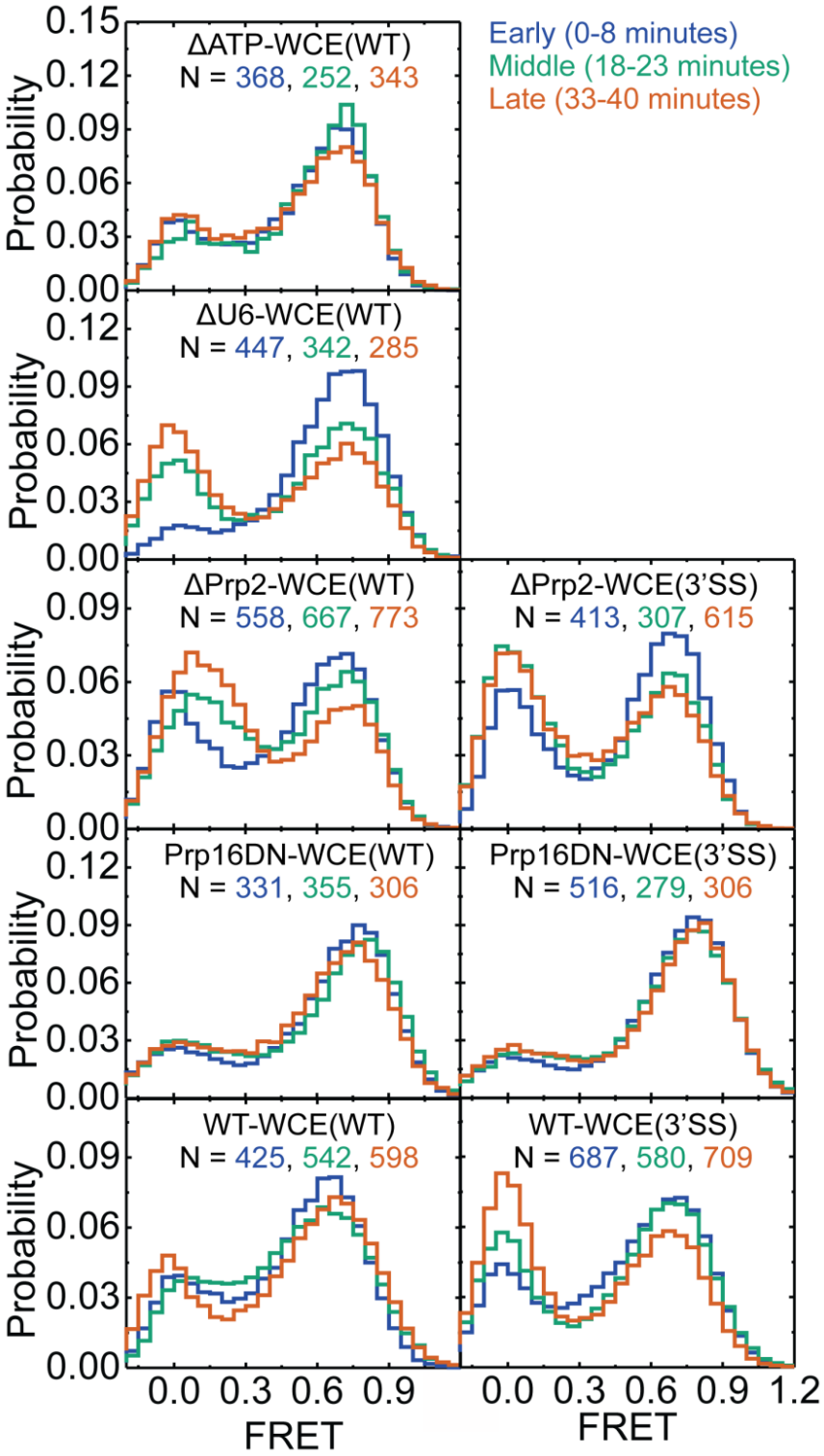
Blanco et al., Supplementary Figure 4



Supplementary Figure 4. Confirmation of blockage and reconstitution of splicing by *in vitro* splicing assays. Denaturing, 7 M urea, 15% (w/v) polyacrylamide gels were scanned with a variable mode Typhoon imager. The intron and intron-lariat products are observed in the Cy3 scan (green) and the mature mRNA product is visualized in the Cy5 scan (red). **(a)** The optimized concentration of D1 required to deplete U6 snRNA (300 nM) was determined by titrating increasing amounts of the oligodeoxynucleotide into the *in vitro* splicing assay. **(b)** Using the previously determined optimal concentration of D1 (300 nM, **a**), extract viability was confirmed through reconstitution with *in vitro* transcribed U6 snRNA. **(c)** Incubation of extract at 37 °C for 40 min completely blocks splicing activity (ΔPrp2 lane). Addition of recombinant Prp2p to ΔPrp2 extract results in reconstitution of splicing, as expected. **(d)** Addition of recombinant dominant mutant Prp16DN to yeast extract stalls splicing after the first chemical step. **(e)** Incubation of WT-WCE with 3'SS mutant substrate stalls splicing after the first step while incubation with a WT substrate **(f)** results in efficient progression through both steps of splicing.

Nature Methods: doi:10.1038/nmeth.3602

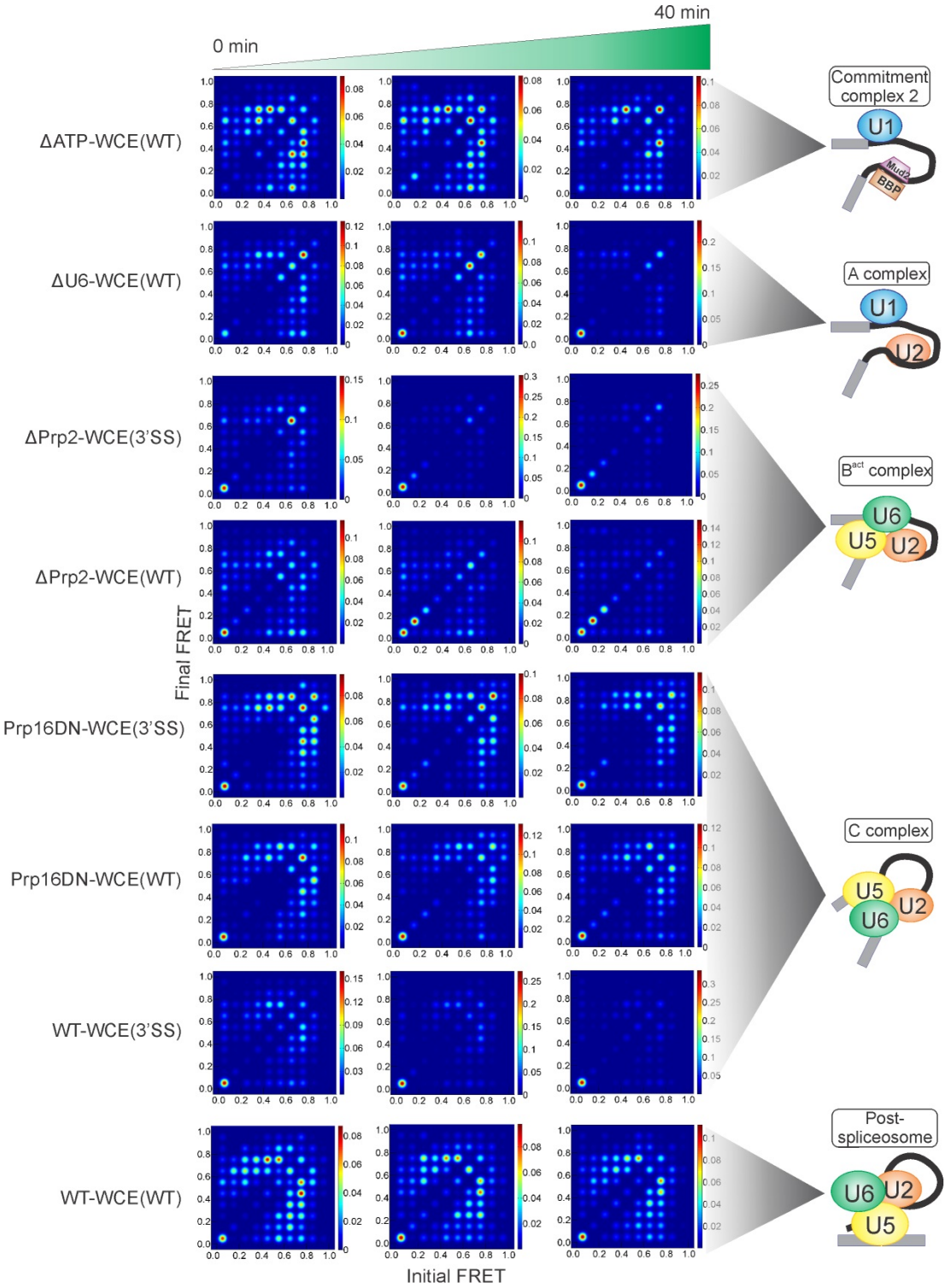
Blanco et al., Supplementary Figure 5



Supplementary Figure 5. FRET probability distribution analysis for each of the 8 experimental conditions over the time course of the smFRET experiments.

Nature Methods: doi:10.1038/nmeth.3602

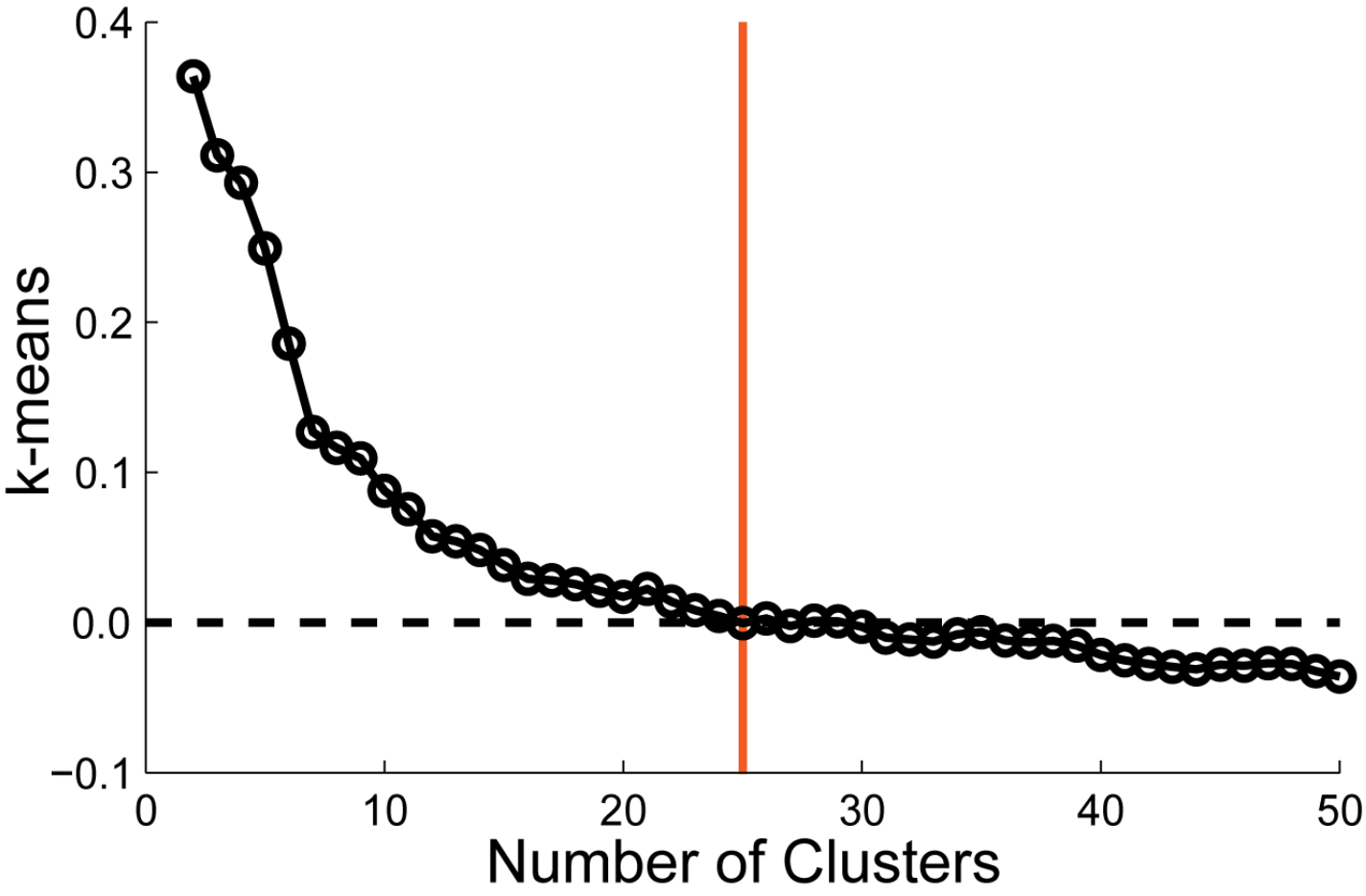
Blanco et al., Supplementary Figure 6



Supplementary Figure 6. Transition Occupancy Density Plots (TODPs) for each of the 8 experimental conditions over the time course of the smFRET experiments depicting the most probable transitions between an initial FRET state (x-axis) and a final FRET state (y-axis).

Nature Methods: doi:10.1038/nmeth.3602

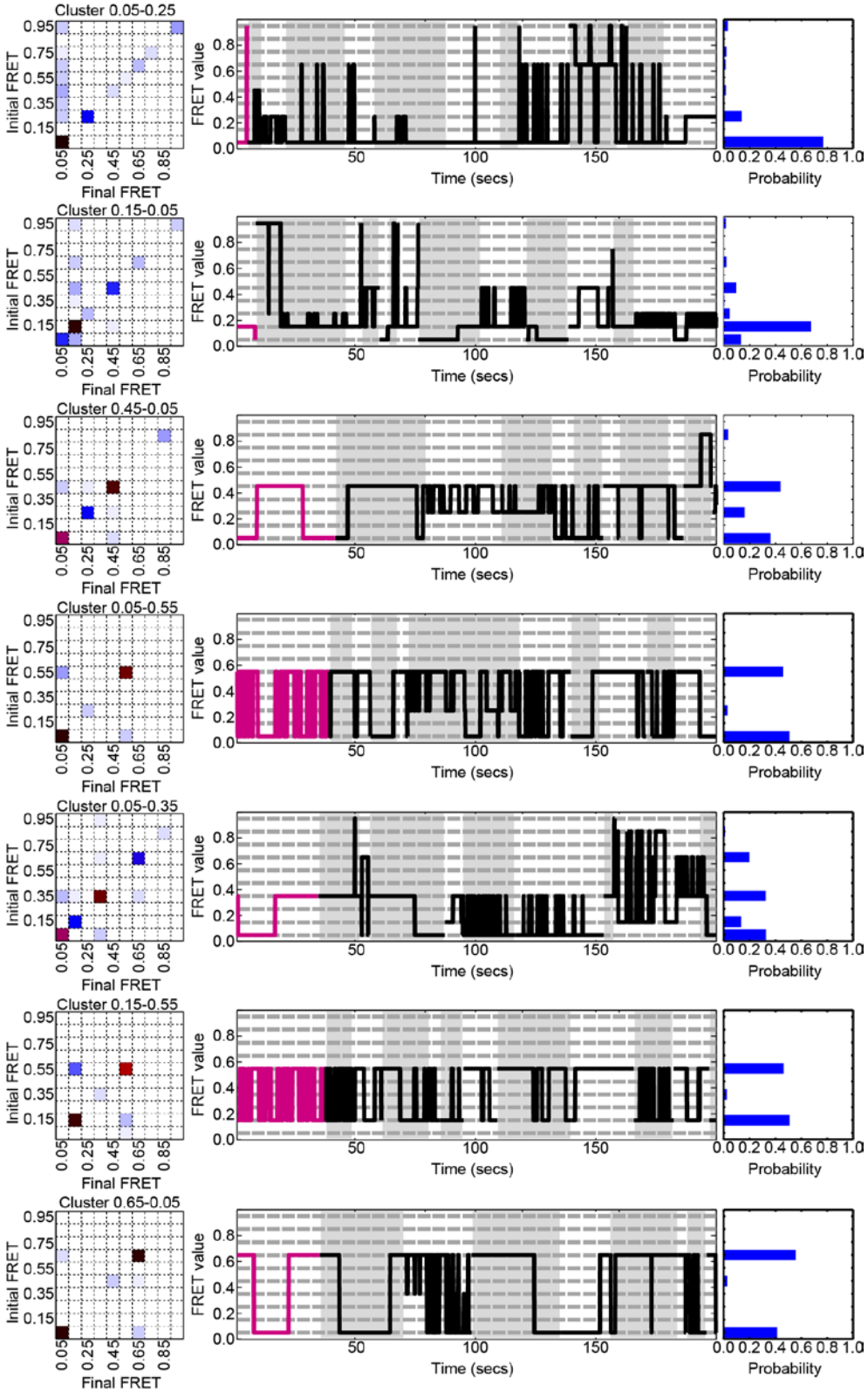
Blanco et al., Supplementary Figure 7



Supplementary Figure 7. 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.

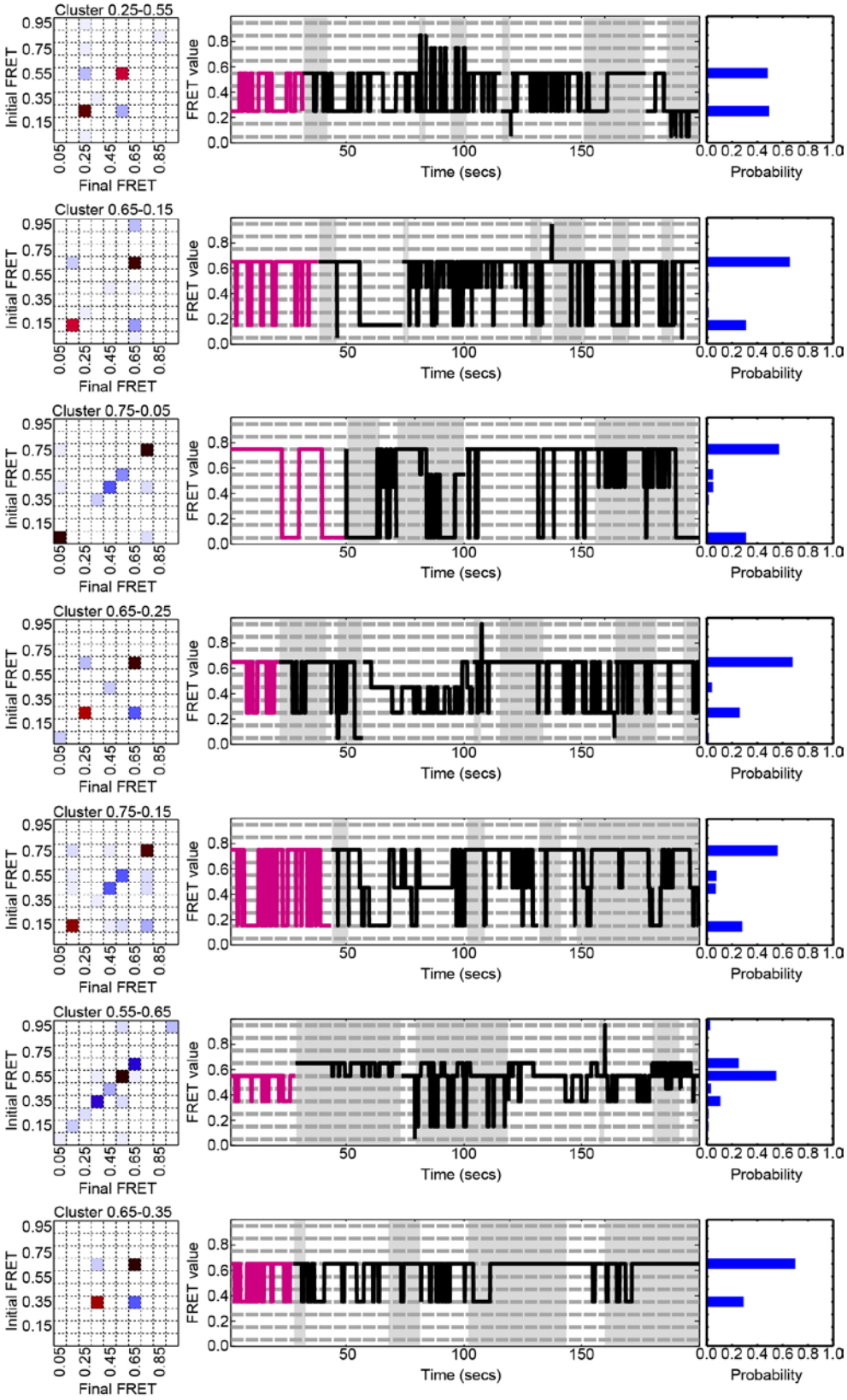
Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 8 - Part 1



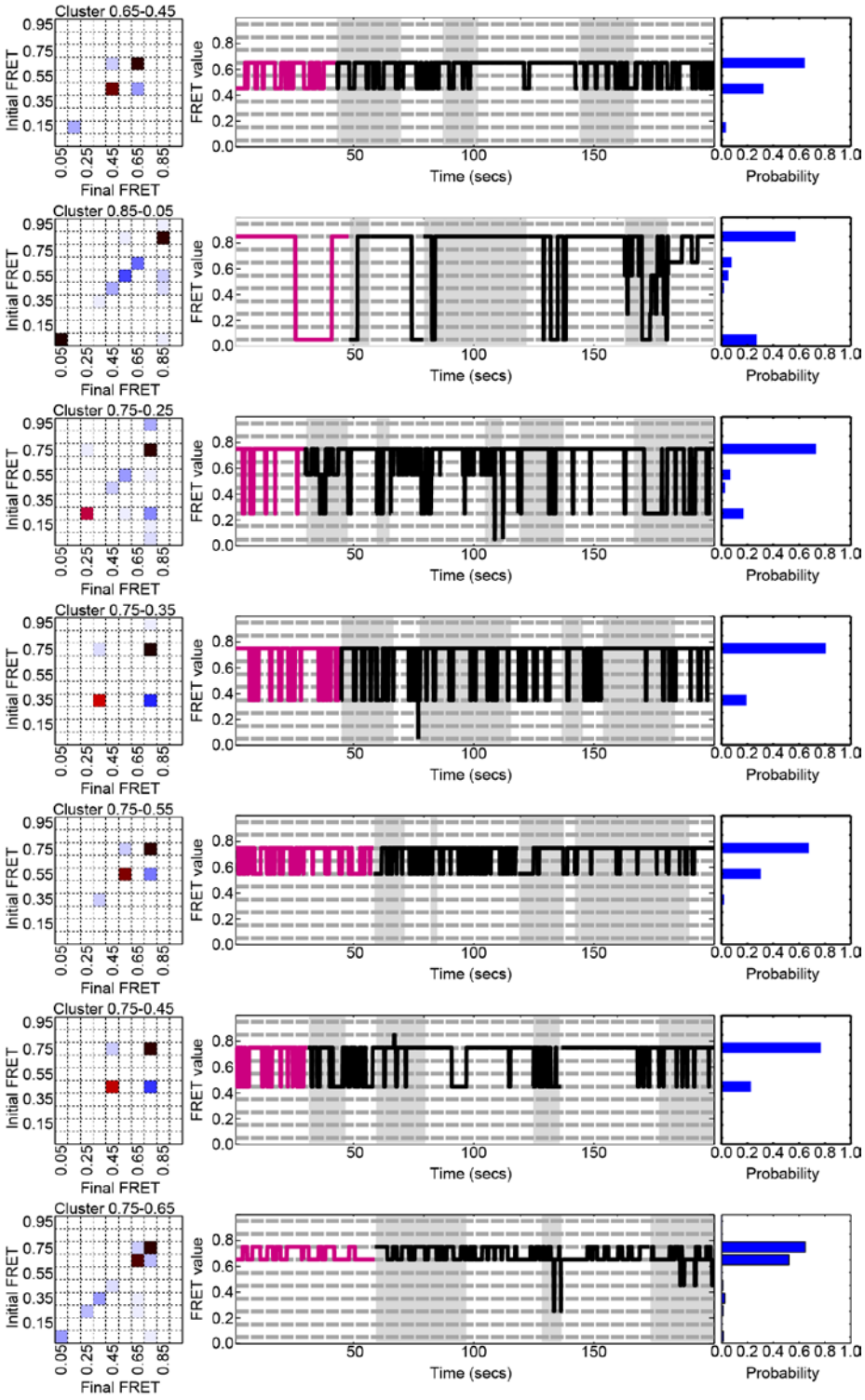
Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 8 - Part 2



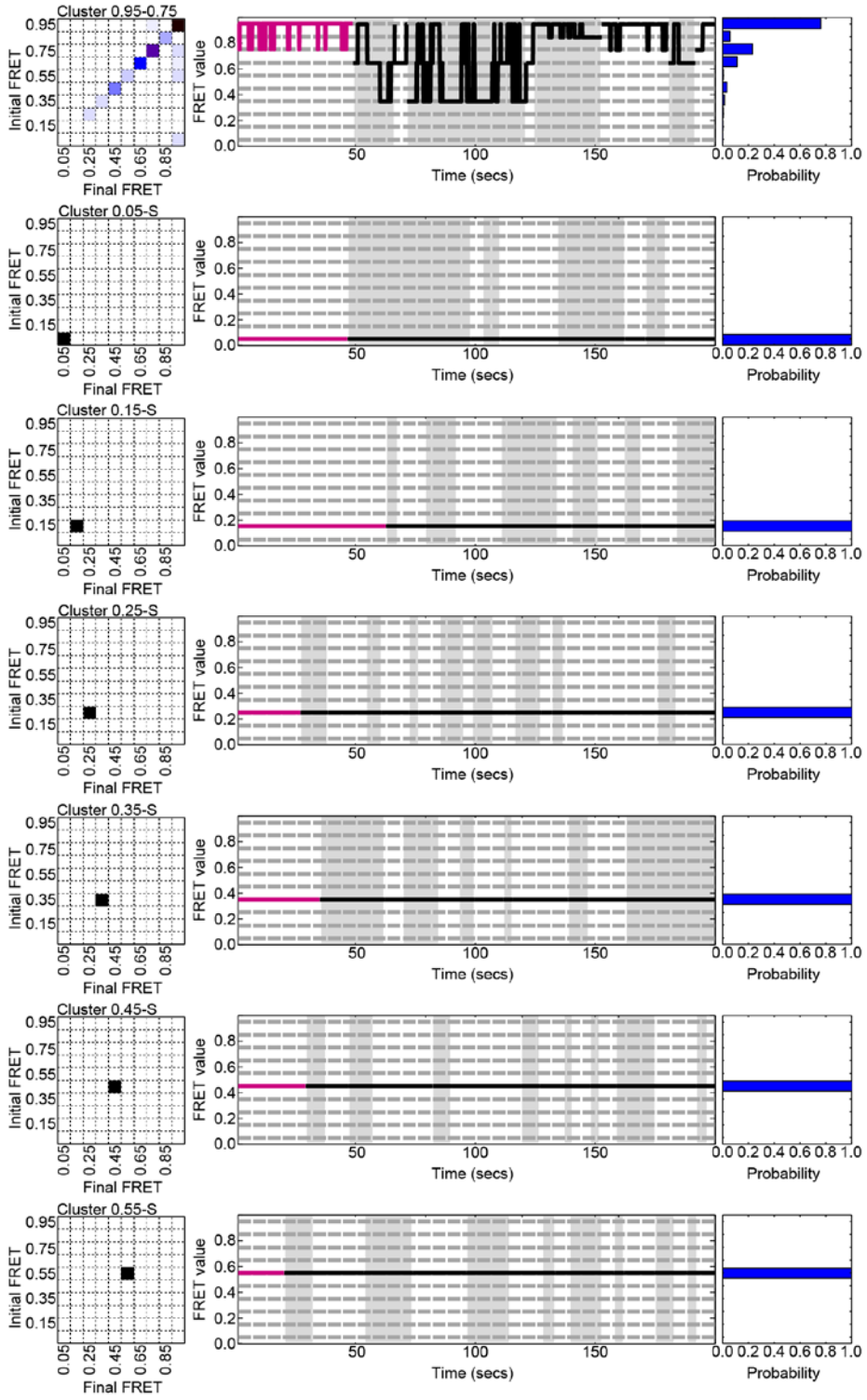
Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 8 - Part 3



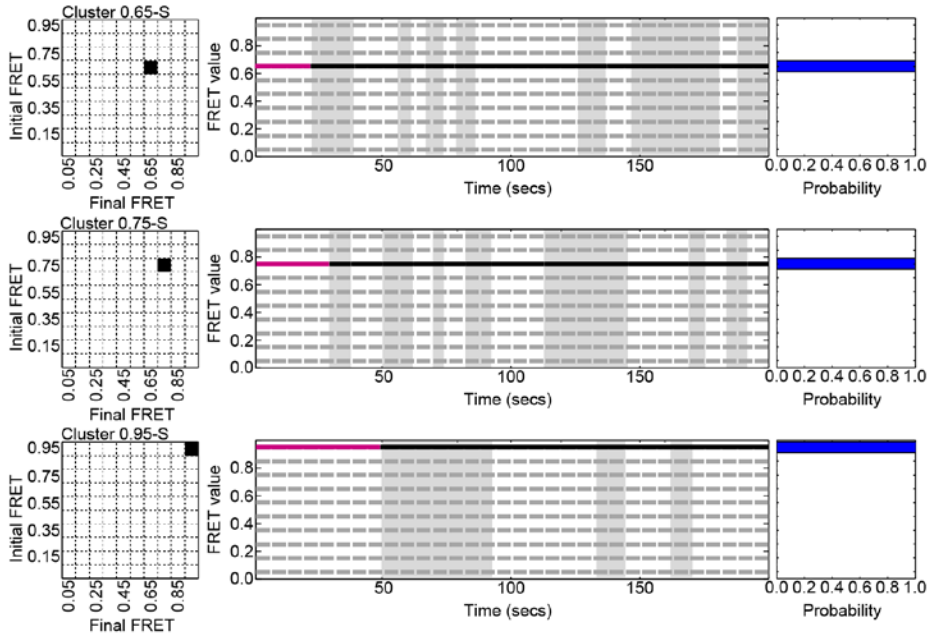
Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 8 - Part 4



Nature Methods: doi:10.1038/nmeth.3602

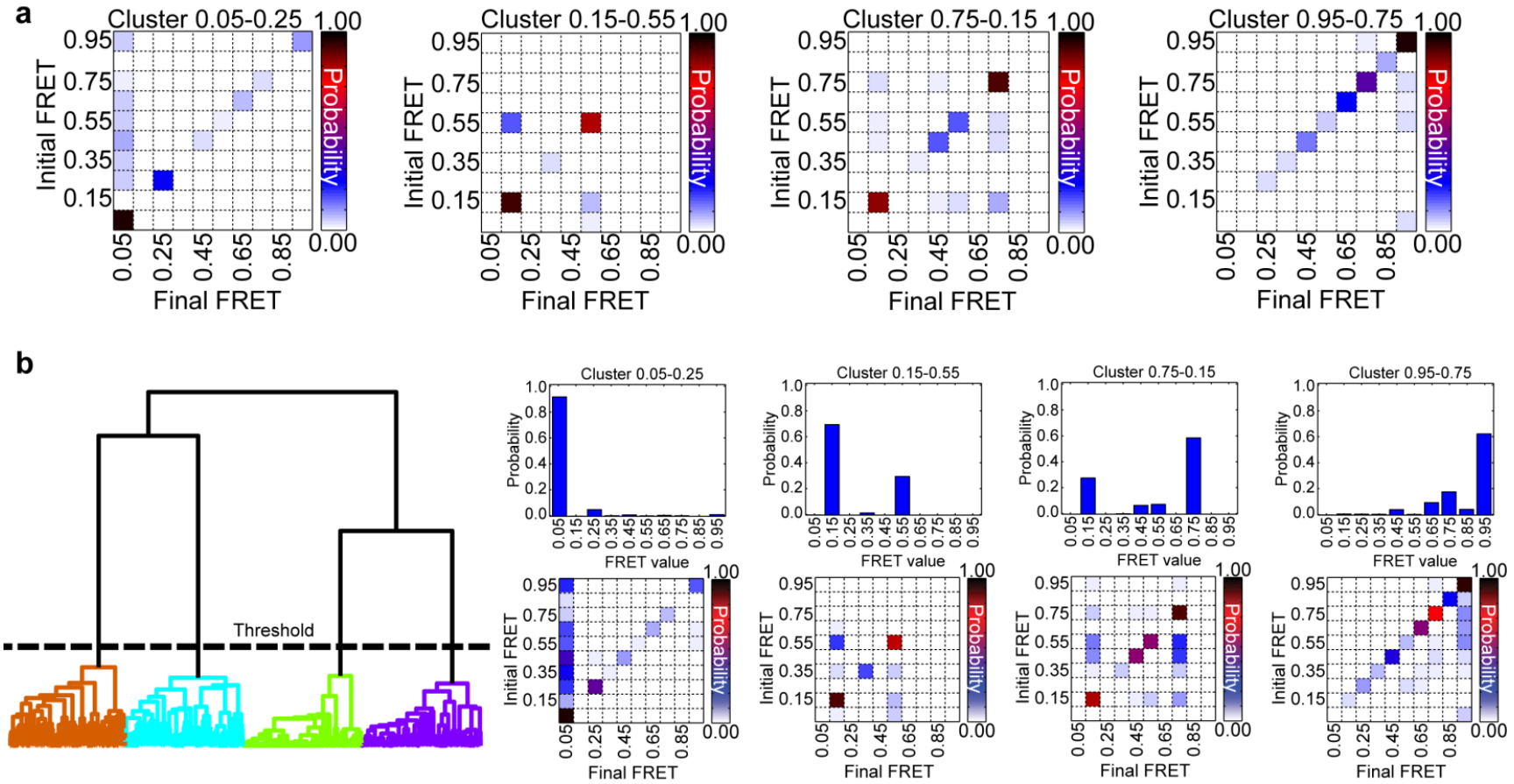
Blanco et al., Supplementary Figure 8 - Part 5



Supplementary Figure 8. Transition probability matrix (left), the longest of the 10 traces whose HMM is most similar to the average HMM of the cluster (magenta) and 200 s of random traces (black, middle), and the probability of FRET states for each dynamic and static cluster (right). Grey and white backgrounds demarcate individual trajectories.

Nature Methods: doi:10.1038/nmeth.3602

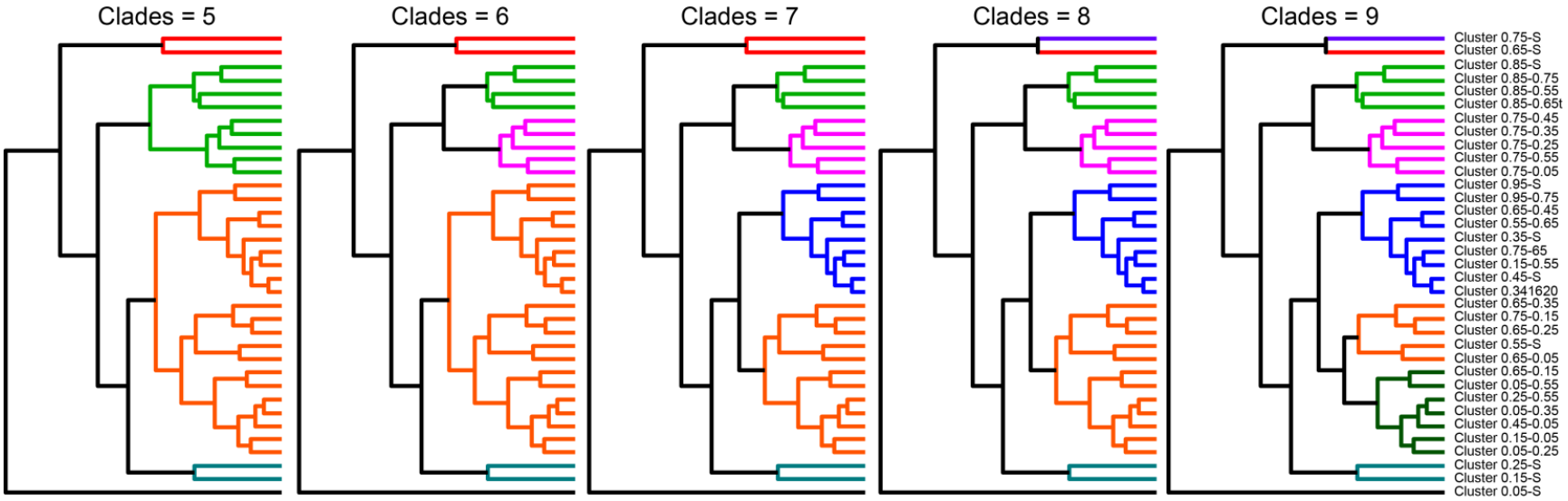
Blanco et al., Supplementary Figure 9



Supplementary Figure 9. Clustering of simulated dataset produced from four of the dynamic clusters representing the large experimental dataset. (a) The four transition probability (TP) matrices from clustering of the full splicing dataset that were used to generate 1500 random traces for each cluster (see online methods). These traces were pooled and used as input for clustering by SiMCAn. **(b)** Hierarchical tree showing the four cluster threshold found by SiMCAn and the four resulting cluster probability histograms and TP matrices.

Nature Methods: doi:10.1038/nmeth.3602

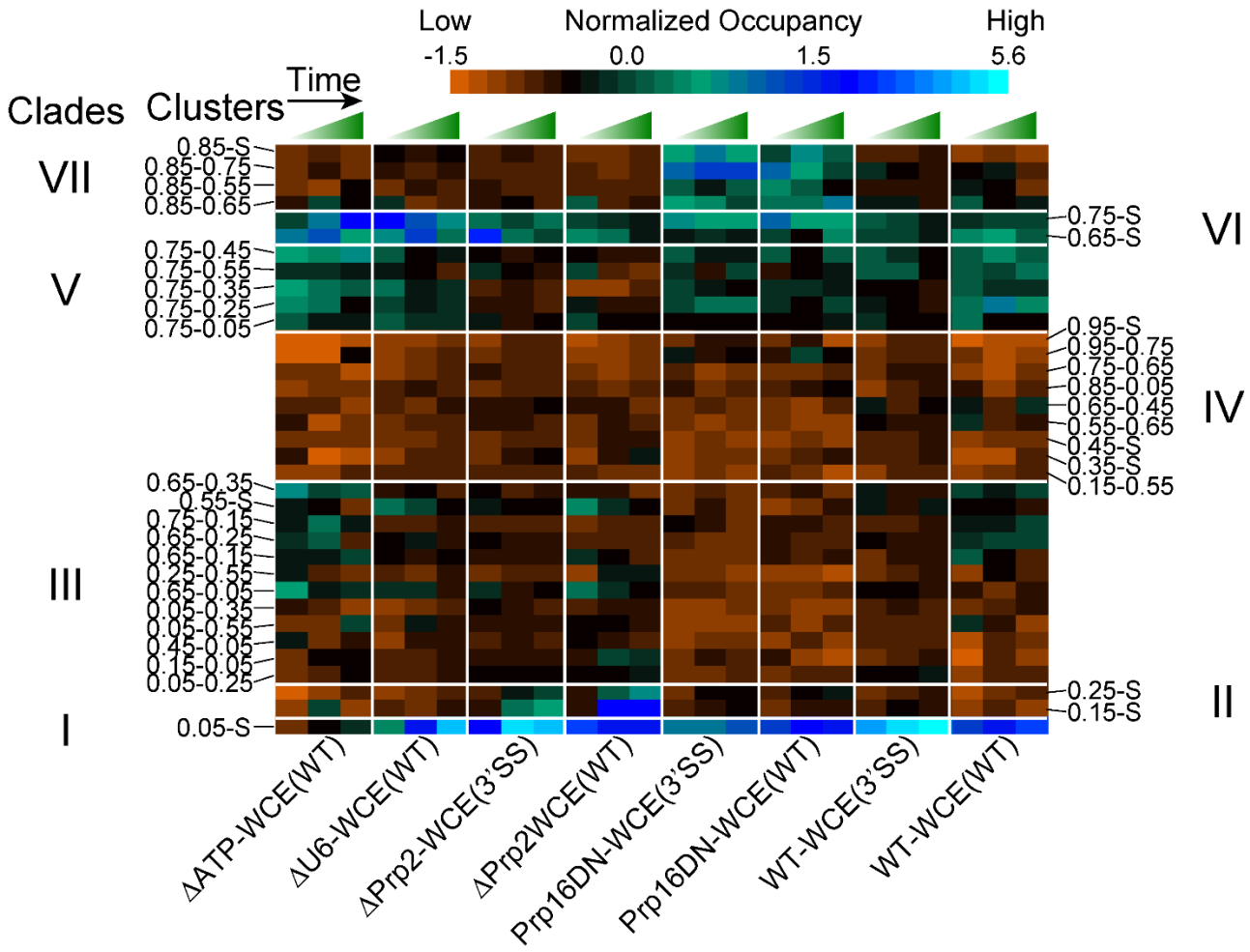
Blanco et al., Supplementary Figure 10



Supplementary Figure 10. Varying the tree cut-off heights upon grouping the cluster occupancy among the 8 experimental conditions leads to distinct numbers of (color-coded) clades of clusters (as indicated on the right).

Nature Methods: doi:10.1038/nmeth.3602

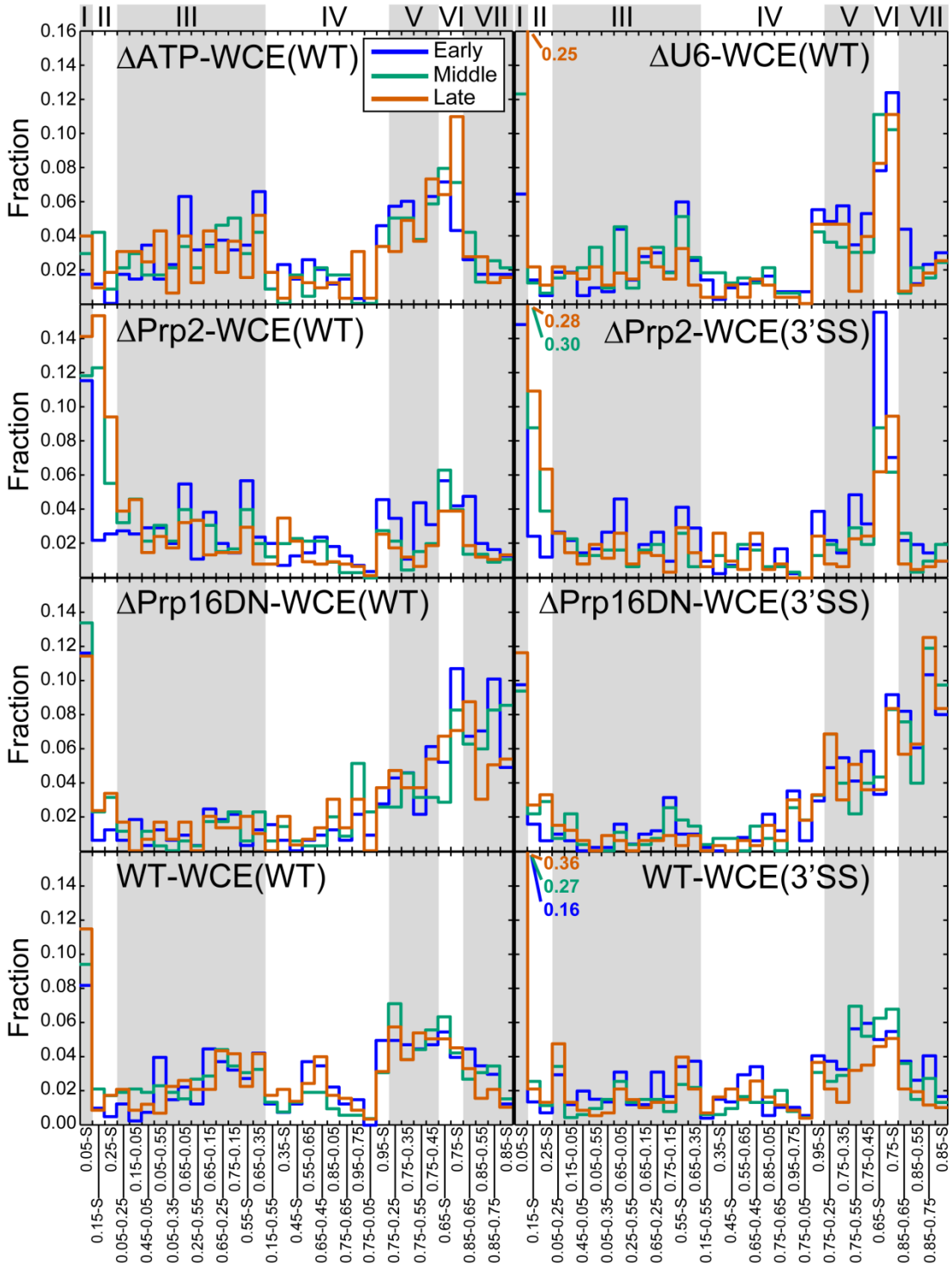
Blanco et al., Supplementary Figure 11



Supplementary Figure 11. Clustering of clusters to identify 'clades' of similar behavior. Heat-map representation of the clustering of clusters for the 8 experimental conditions.

Nature Methods: doi:10.1038/nmeth.3602

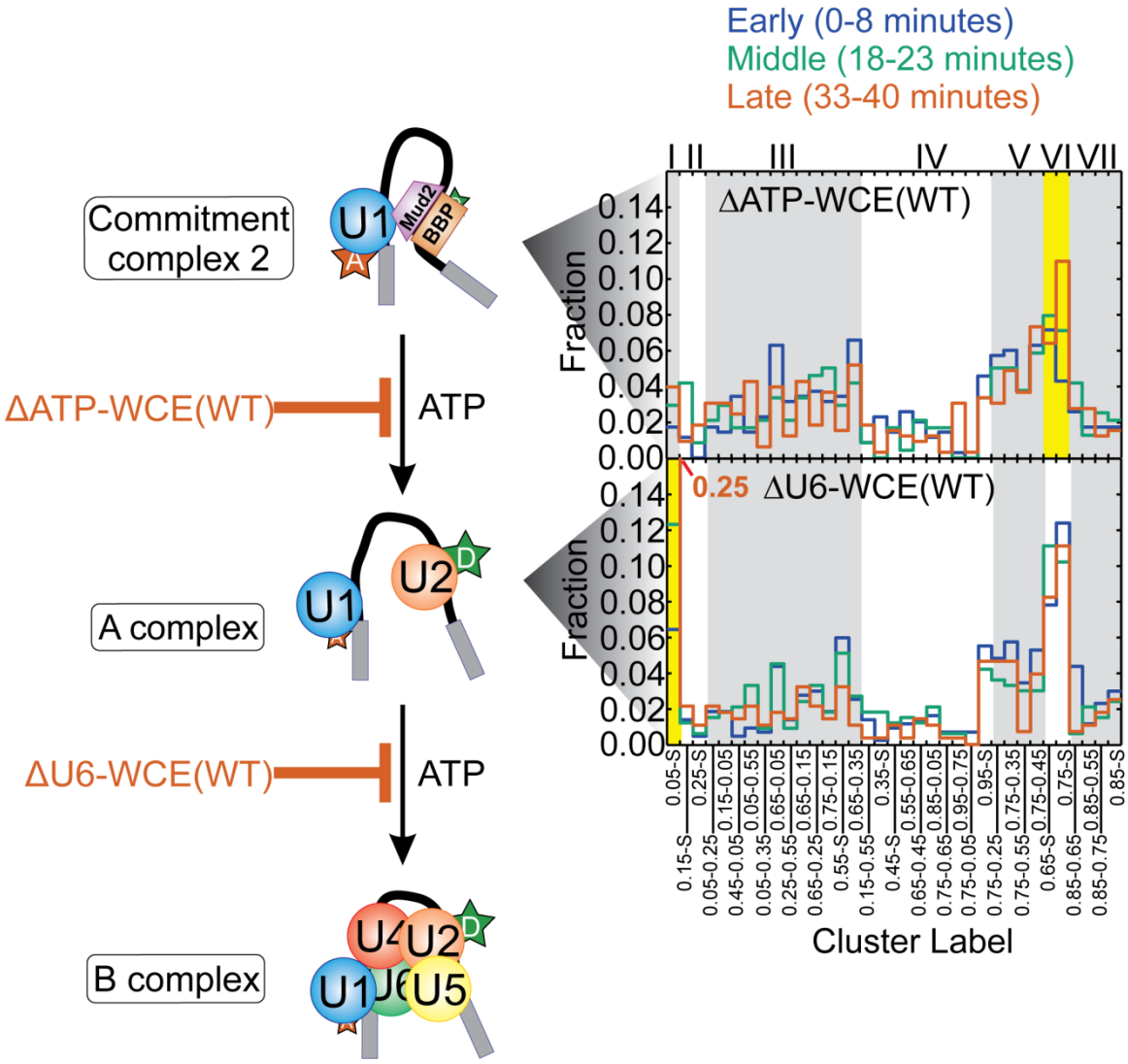
Blanco et al., Supplementary Figure 12



Supplementary Figure 12. Histogram showing the raw fraction of molecules occupying each cluster of the 8 experimental conditions. Alternating gray and white backgrounds demarcate the clusters (bottom) comprising each of the 7 clades (top).

Nature Methods: doi:10.1038/nmeth.3602

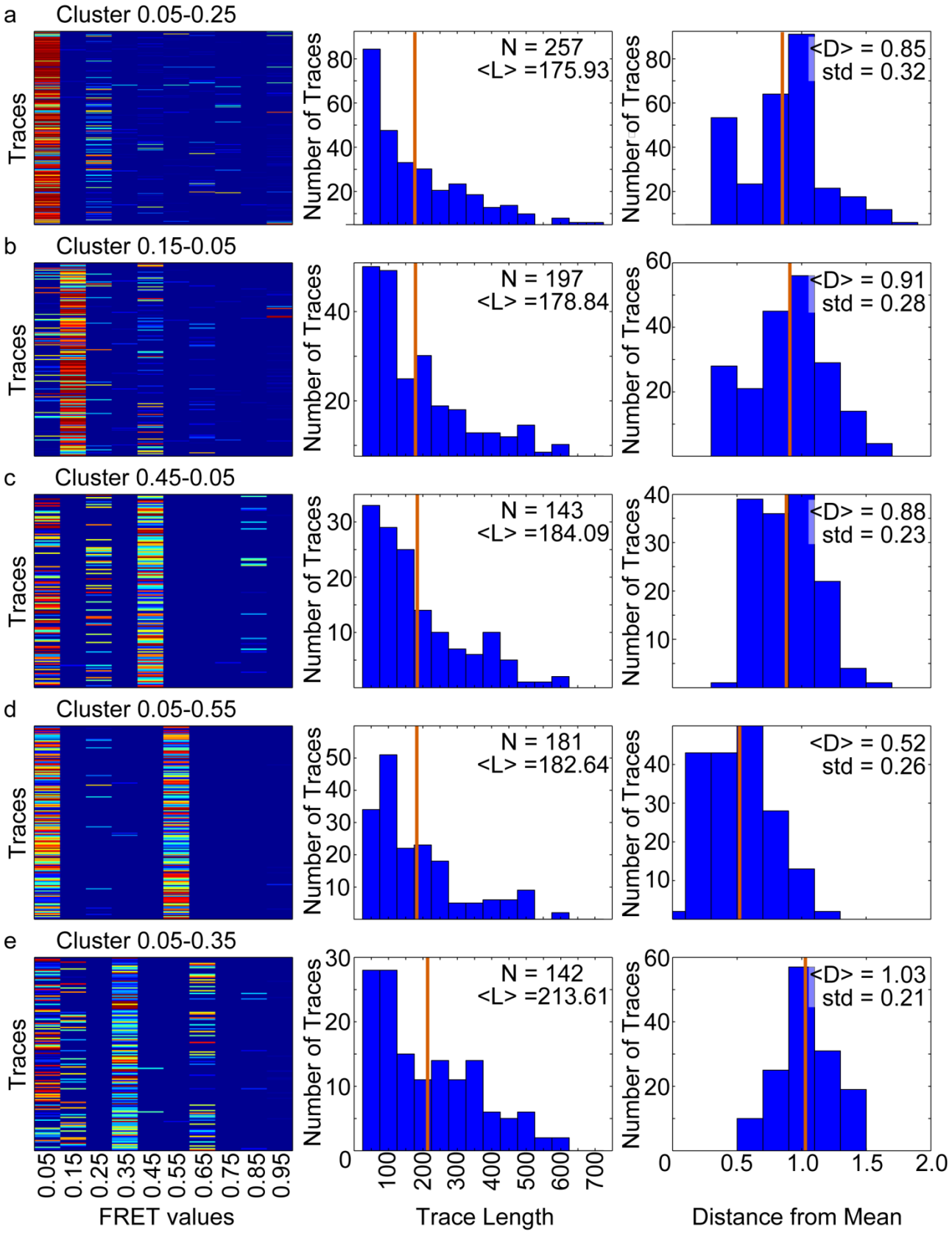
Blanco et al., Supplementary Figure 13



Supplementary Figure 13. Histogram showing the raw fraction of molecules occupying each cluster for the early splicing block conditions. Alternating gray and white backgrounds demarcate the clusters (bottom) comprising each of the 7 clades (top). Clusters of significant occupancy within a specified condition are highlighted in yellow.

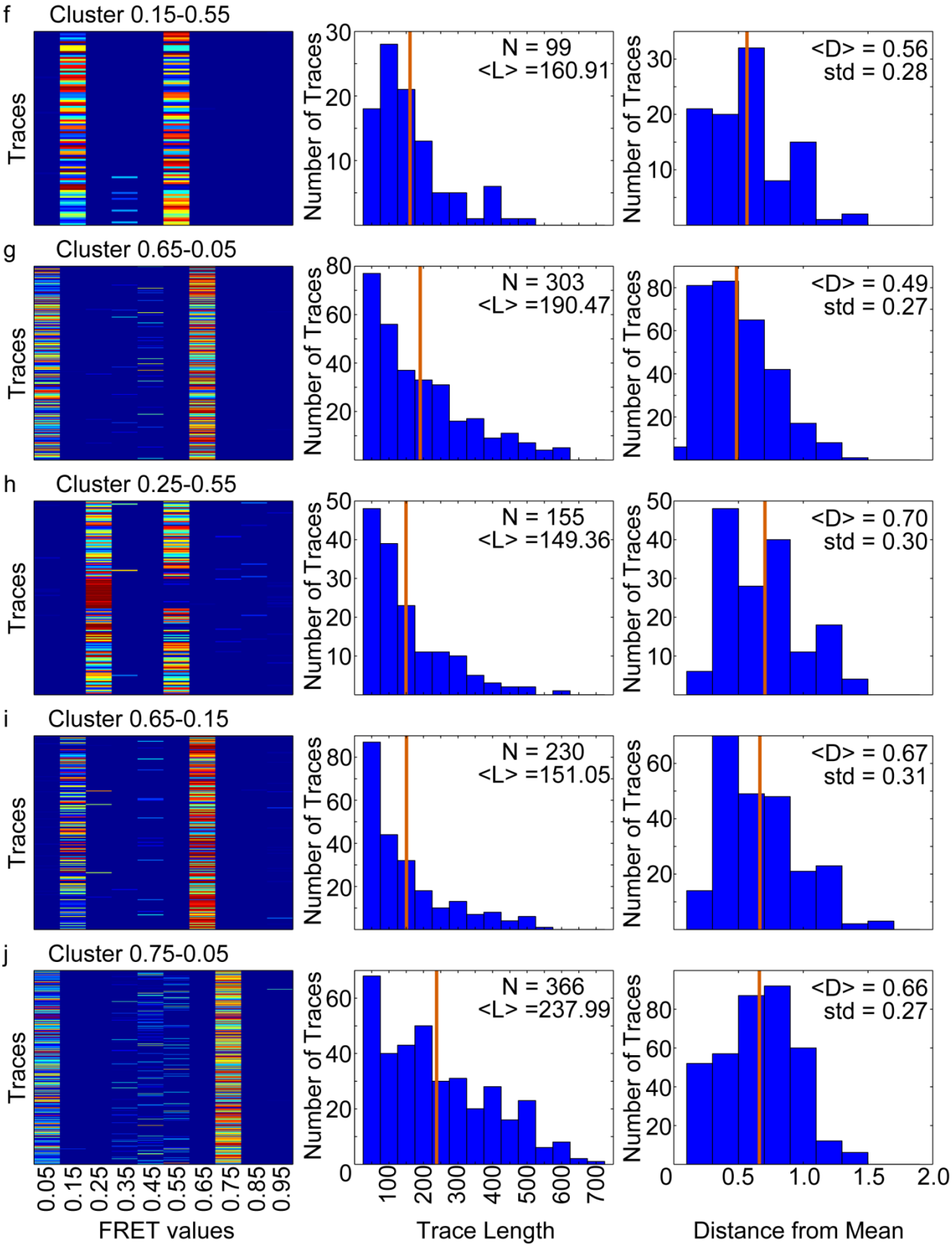
Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 14 - Part 1



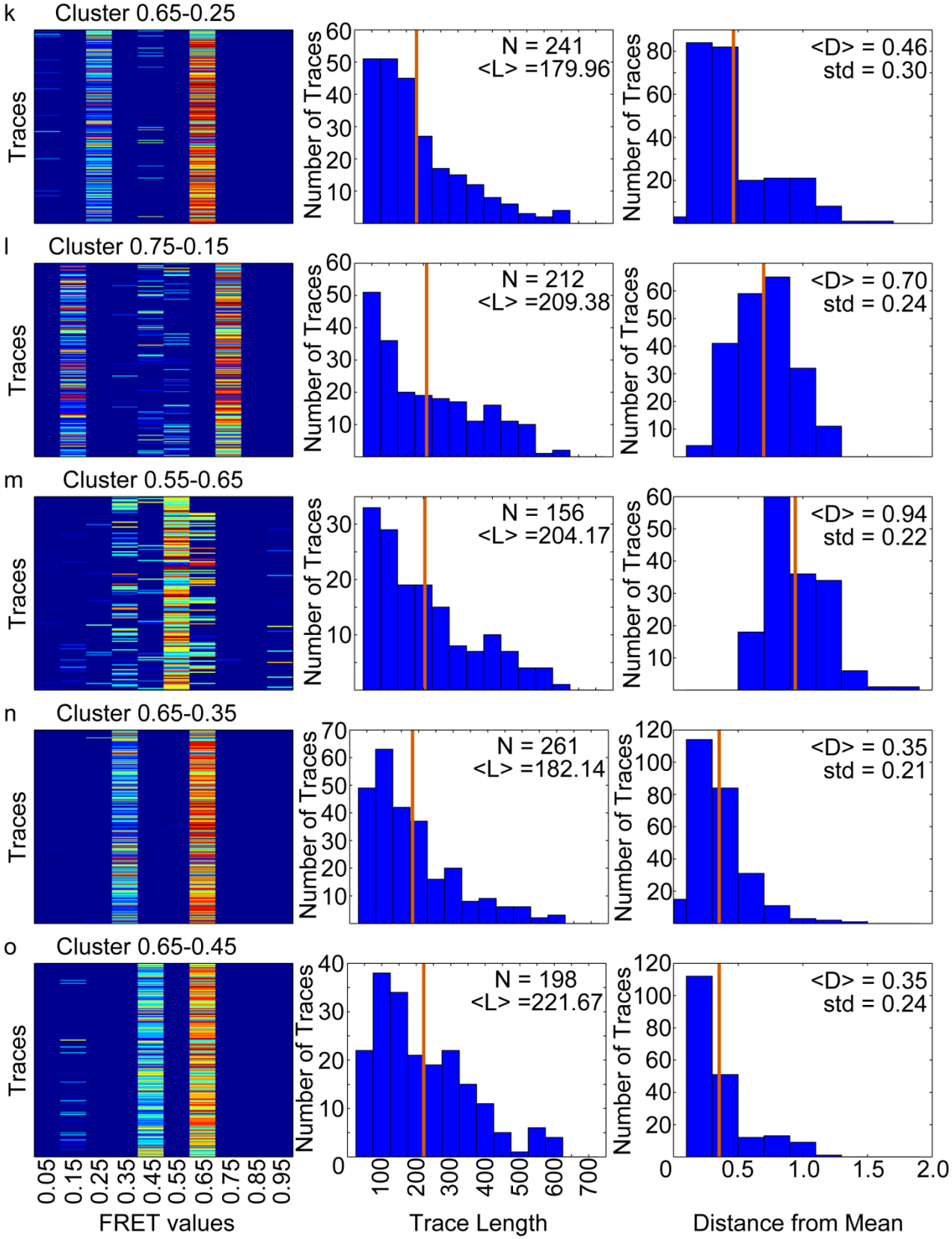
Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 14 - Part 2



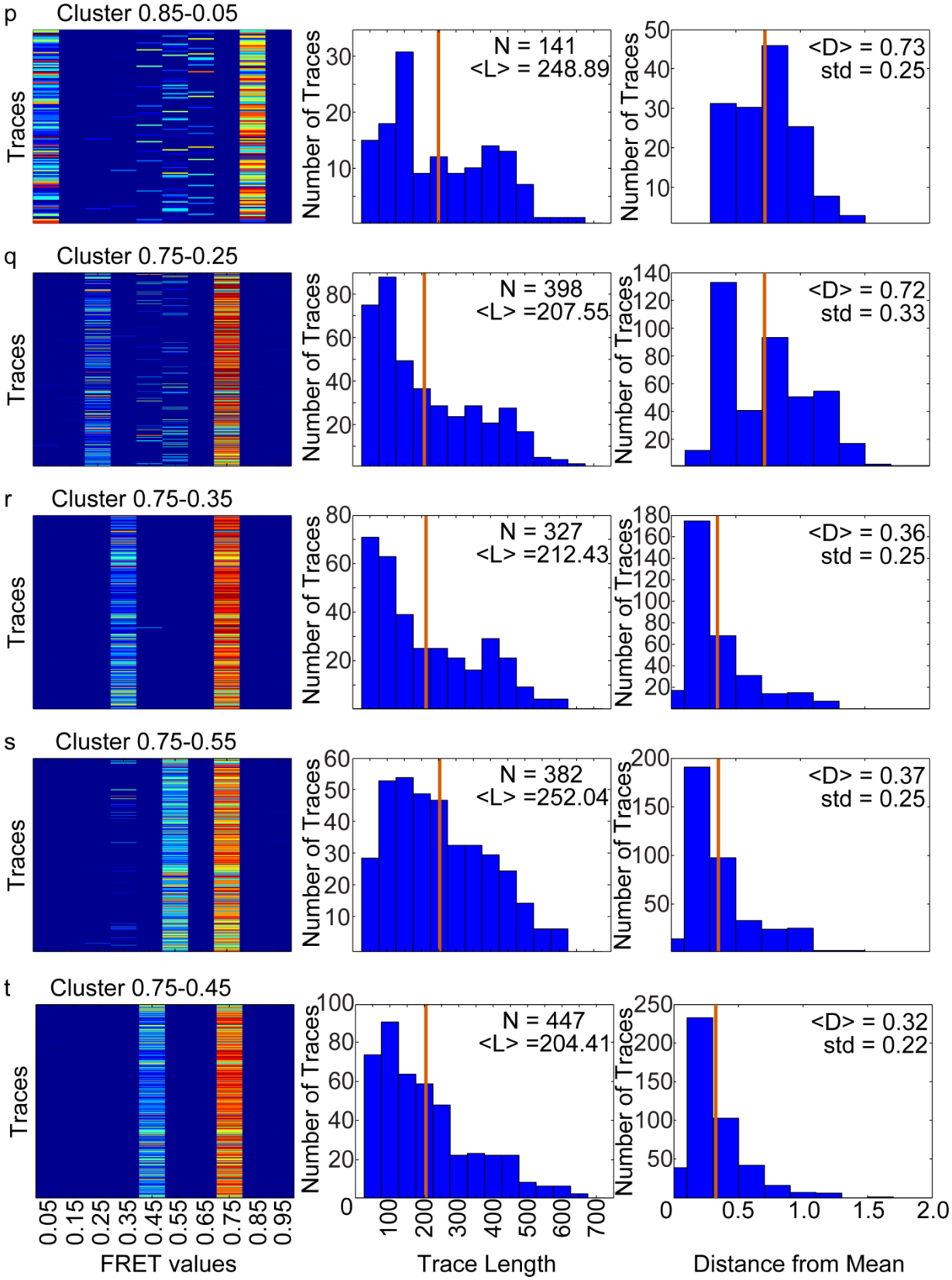
Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 14 - Part 3



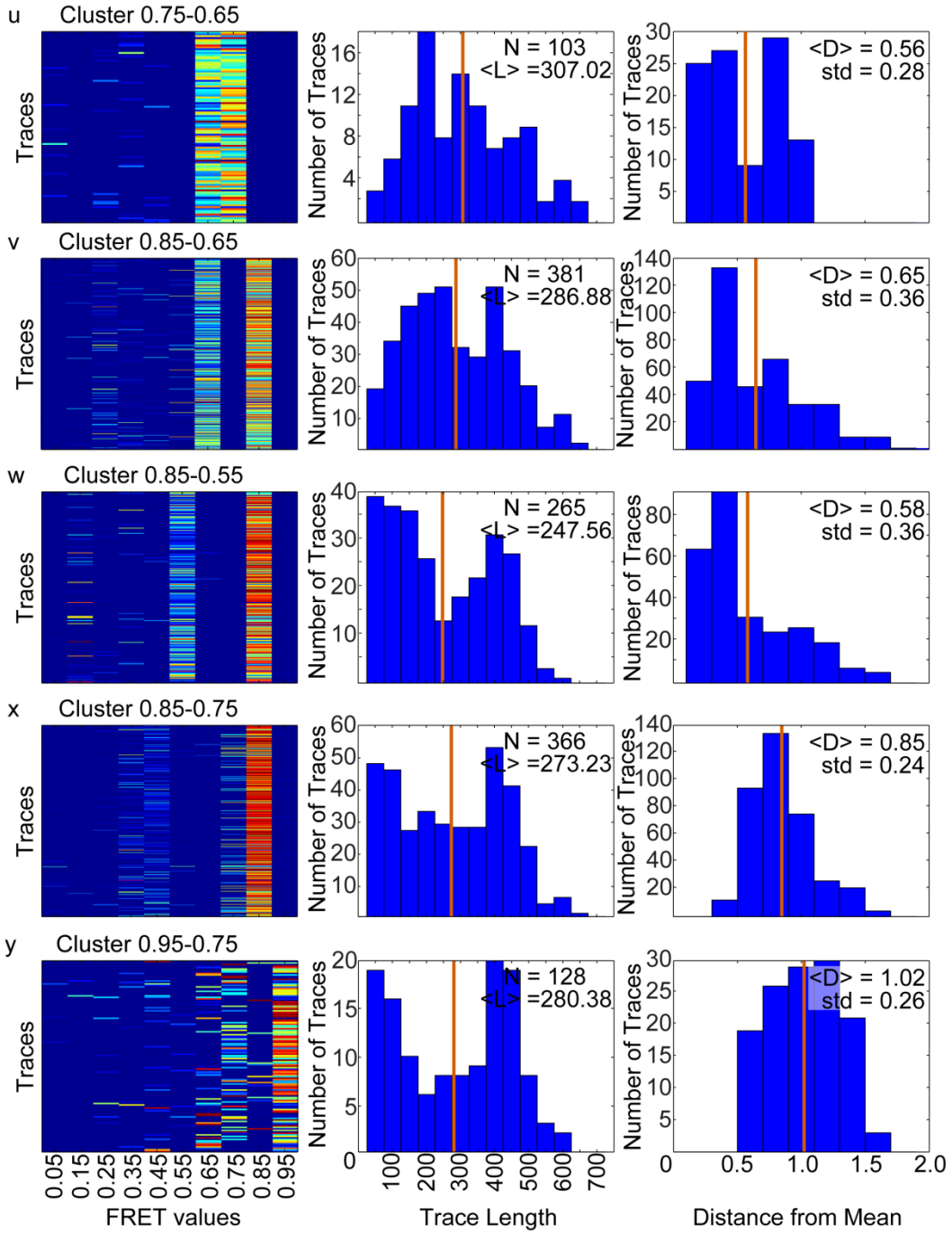
Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 14 - Part 4



Nature Methods: doi:10.1038/nmeth.3602

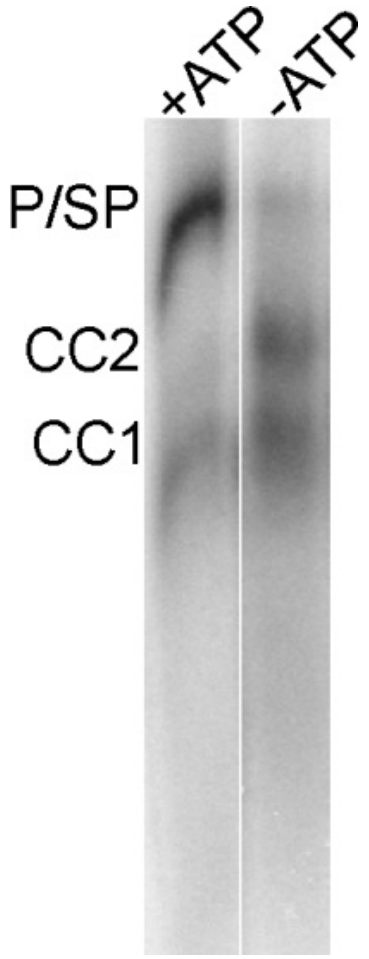
Blanco et al., Supplementary Figure 14 - Part 5



Supplementary Figure 14. Statistical analysis of all 35 (25 dynamic, 10 static) clusters. (a-y) The left panel depicts every trace that contributes to a cluster with the heat map indicating the occupancy at the state (blue = 0, red = 1). The middle panel indicates the distribution of trace length in each cluster. The number of molecules in each cluster (N) and the average trace length (<L>, red line) are indicated in the top right corner. The right panel plots the distance of each trace's HMM from the mean HMM of the cluster. The average distance (<D>, red line) and standard deviation (std) are indicated.

Nature Methods: doi:10.1038/nmeth.3602

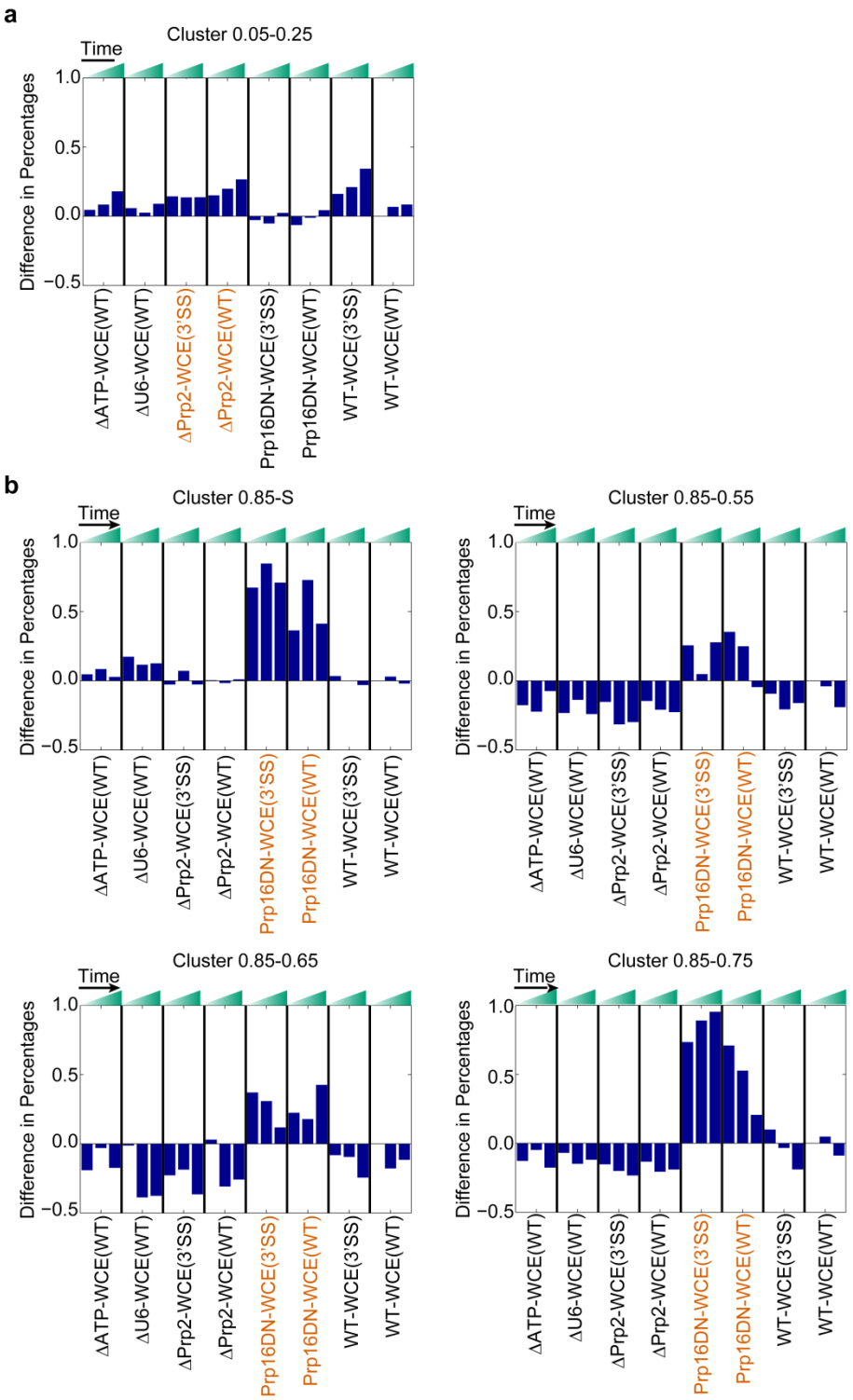
Blanco et al., Supplementary Figure 15



Supplementary Figure 15. Native gel analysis of commitment complex formation upon Ubc4 in BJ2168 extract.

Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 16



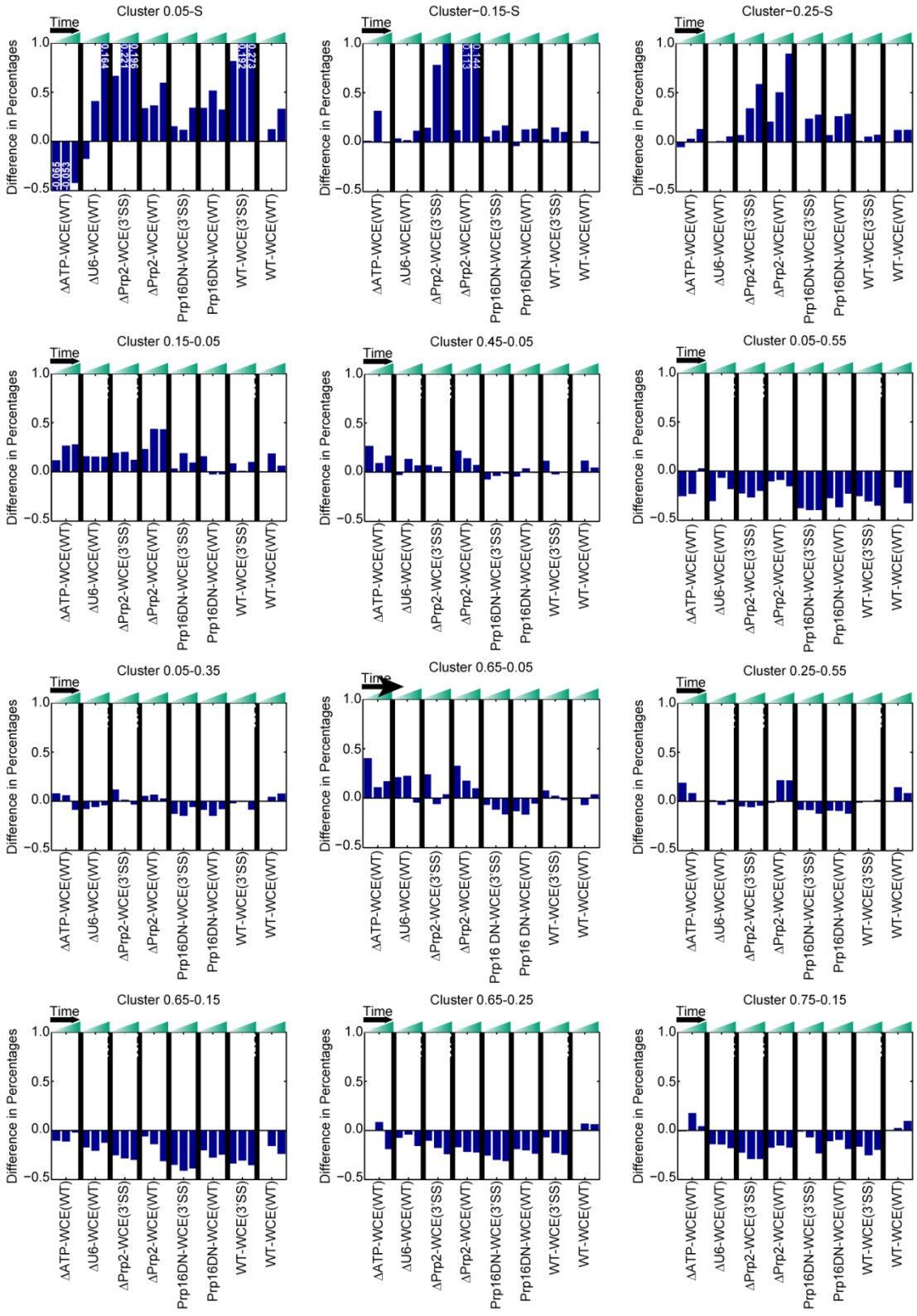
Supplementary Figure 16. The occupancy of clusters within each of the 8 experimental conditions compared to that of WT-WCE(WT). Each occupancy value for every condition is subtracted from the occupancy of the cluster in the WT-WCE(WT) early condition. **(a)** Cluster

Nature Methods: doi:10.1038/nmeth.3602

enriched in the ΔPrp2-WCE condition (red). **(b)** Clusters enriched in the Prp16DN-WCE condition (red).

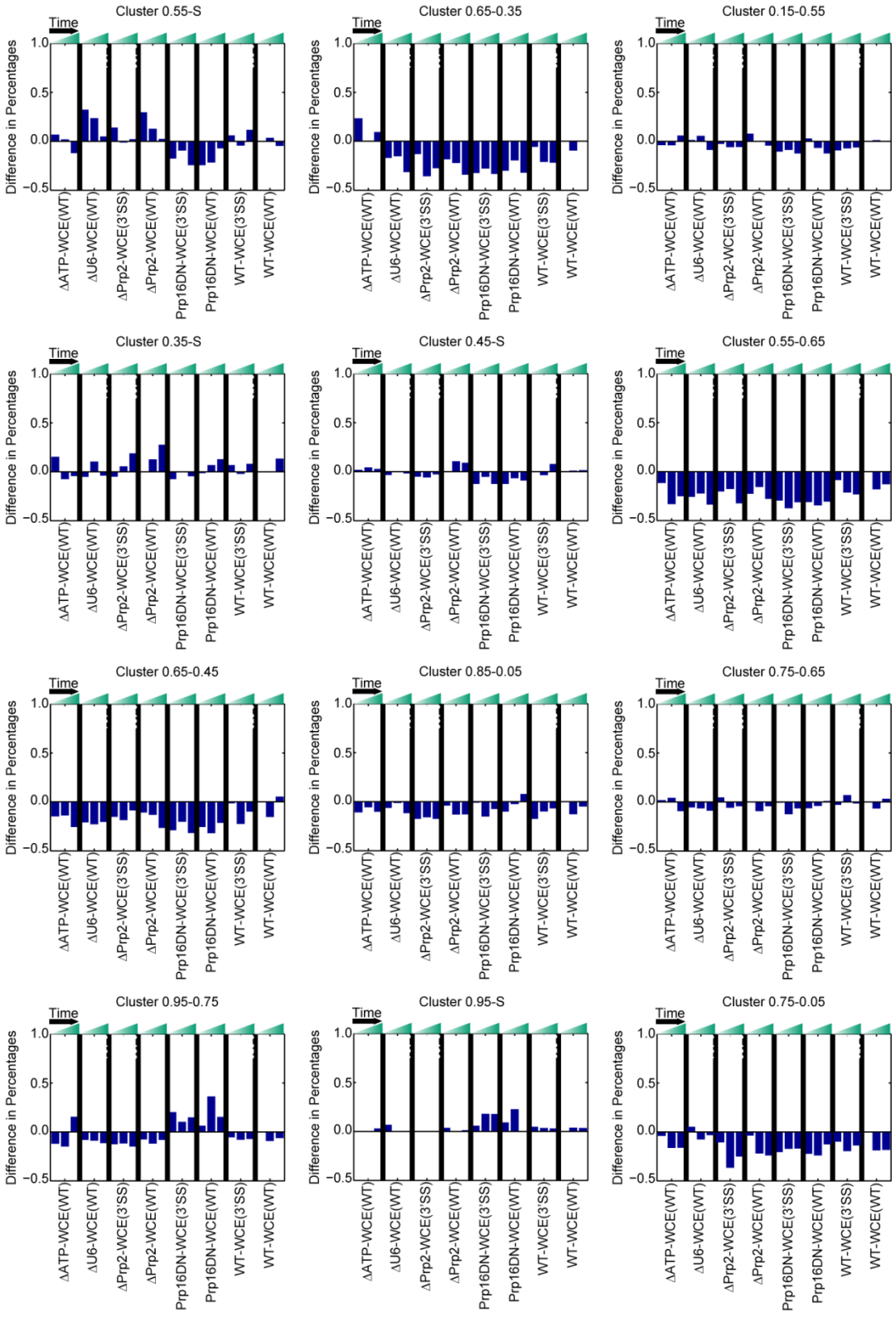
Blanco et al., Supplementary Figure 17 - Part 1

Nature Methods: doi:10.1038/nmeth.3602



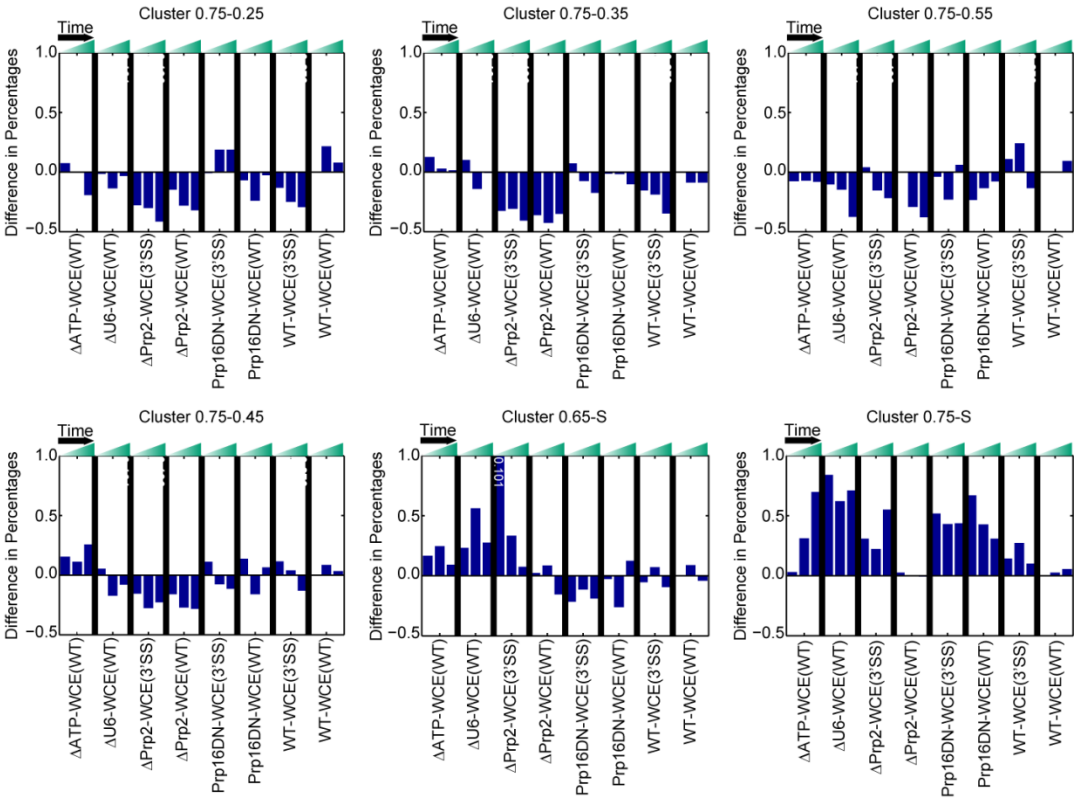
Blanco et al., Supplementary Figure 17 - Part 2

Nature Methods: doi:10.1038/nmeth.3602



Blanco et al., Supplementary Figure 17 - Part 3

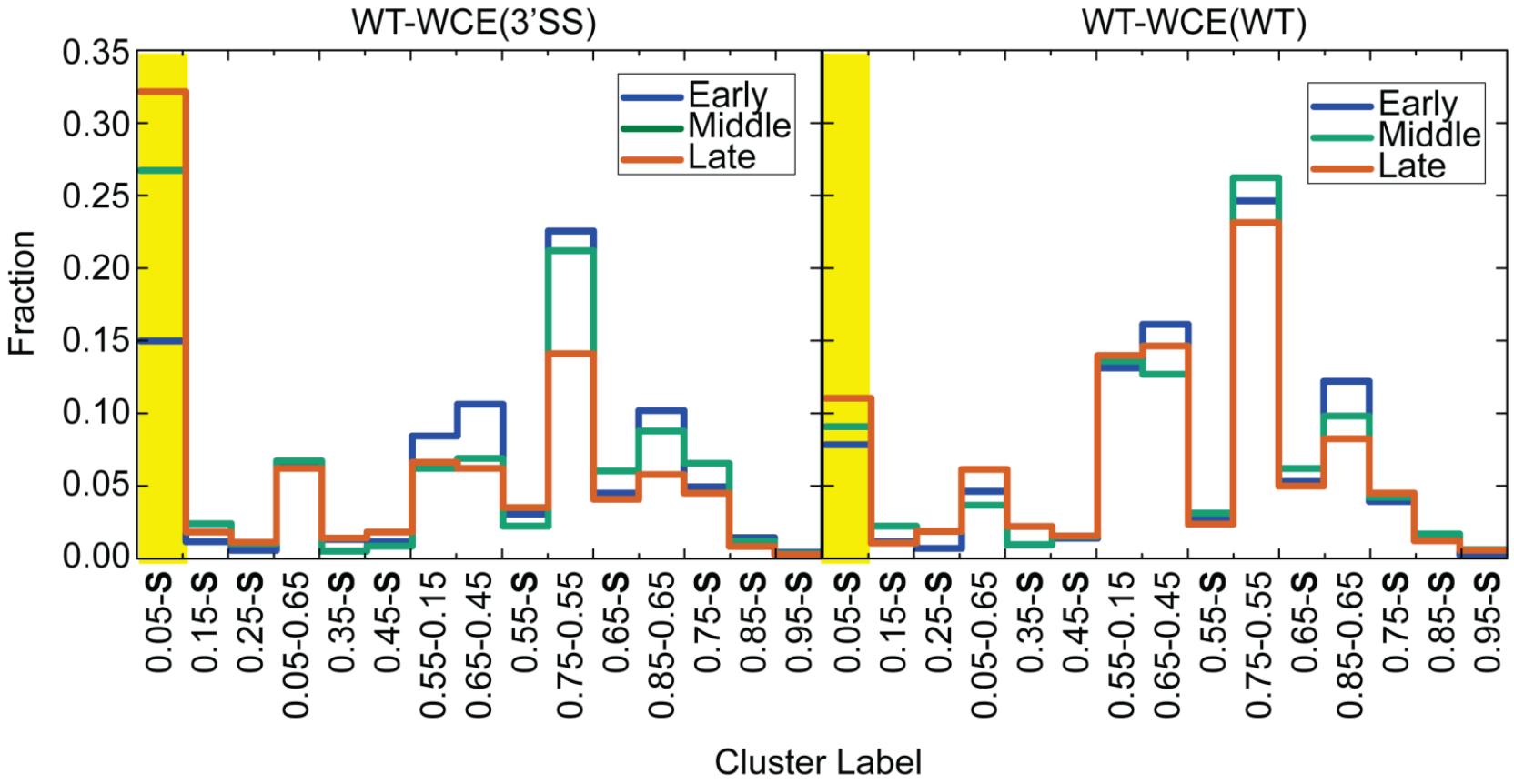
Nature Methods: doi:10.1038/nmeth.3602



Supplementary Figure 17. The occupancy of clusters within each of the 8 experimental conditions compared to that of WT-WCE(WT), as in Supplementary figure 13, for the remaining 30 clusters.

Nature Methods: doi:10.1038/nmeth.3602

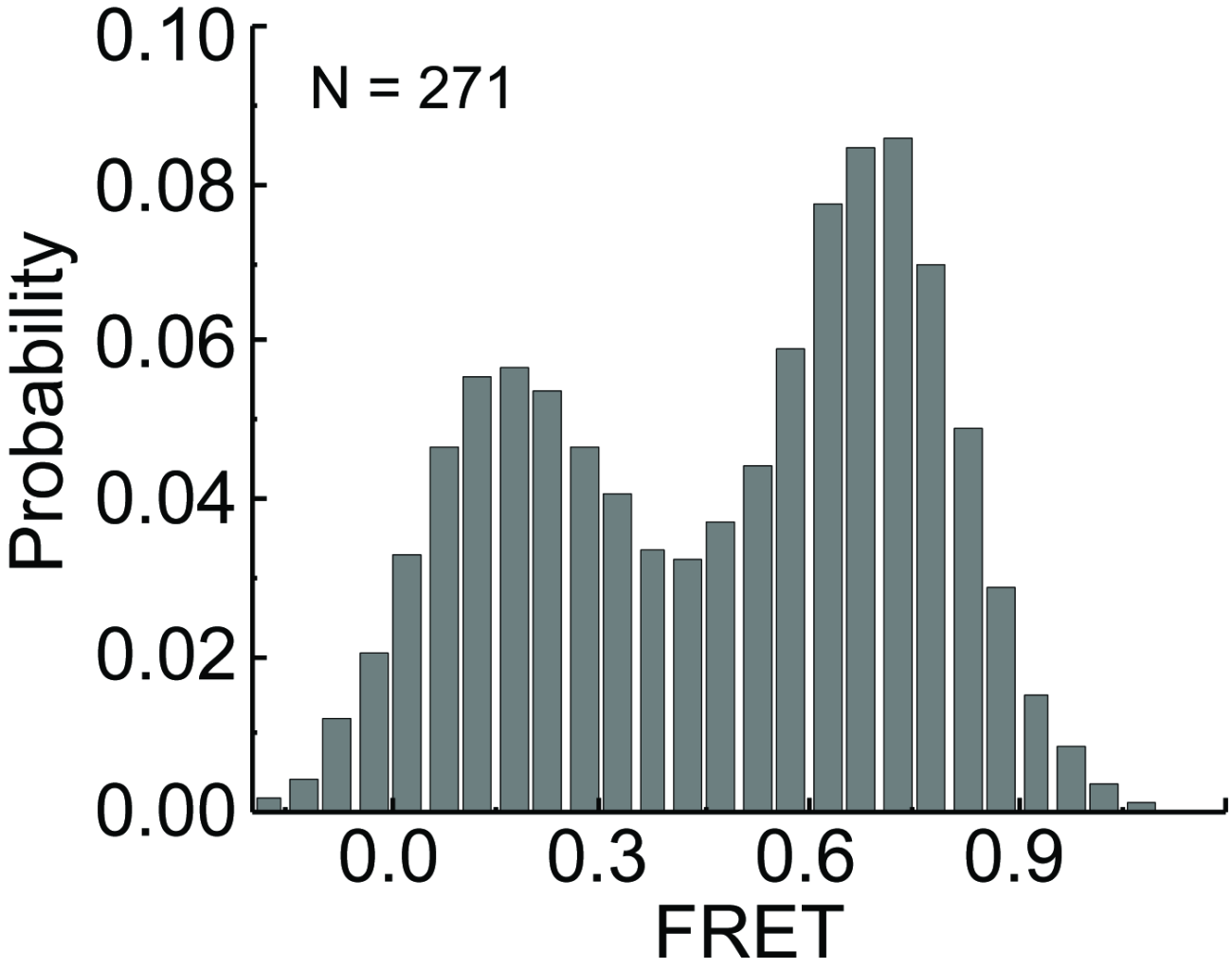
Blanco et al., Supplementary Figure 18



Supplementary Figure 18. Clustering of molecules belonging to the WT-WCE(3'SS) and WT-WCE(WT) conditions reveals enrichment of the 0.05-S cluster with the mutant substrate. Clustering resulted in 5 dynamic and 10 static clusters.

Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Figure 19



Supplementary Figure 19. smFRET analysis of fluorophore-labeled Ubc4 containing the 3'SS mutation in splicing buffer in the absence of any spliceosomal components.

Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Table 1

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

Supplementary Table 1. Substrate and extract used to form the splicing complexes in each of the experimental conditions employed in this study.

Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Table 2

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

Supplementary Table 2. Statistical analysis of each of the 35 clusters.

Nature Methods: doi:10.1038/nmeth.3602

Blanco et al., Supplementary Table 3

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

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

Nature Methods: doi:10.1038/nmeth.3602

Supplementary Note 1: General notes about SiMCAn methodology and spliceosome heterogeneity

Although we show through splicing assays that the indicated mutation results in enrichment of a particular splicing complex, there is still spliceosome heterogeneity present on the slide surface (due to incomplete assembly/catalysis, variation in post-translational modification, etc.). However, a significant fraction of molecules do exhibit behavior associated with the indicated splicing complex. Additionally, any heterogeneity in spliceosome content will result in a change in FRET behavior from the dominant complex/condition and thus will be identified by SiMCAn through the assignment of clades. This second layer of hierarchical clustering allows for the identification of multiple clusters that correlate based on their occupancies in different experimental conditions. This helps re-bin behaviors that are characterized by slightly different FRET states and allows for the use of kinetics as an additional parameter to resolve FRET states with similar FRET efficiencies. Such an approach likely allows for the grouping of molecules with subtle differences in protein composition or post-translational modification, but that have progressed to the same stage of splicing assembly or catalysis. Furthermore, this second round of clustering should account for any day-to-day variability in fluorescent background, which may result in a small change in FRET signal and HMM fitting. An example of this is seen in the ΔPrp2-WCE(WT,3'SS) datasets in which the 0.15-**S** and 0.25-**S** clusters are both enriched. The pre-mRNA likely adopts an intermediate conformation that, due to day-to-day variability or slight differences in protein composition, is binned into either the 0.15-**S** or 0.25-**S** clusters.

We use 10 bins in our analysis due to the noise-limited FRET resolution of our single-molecule experimentation. Ideally, one would want to choose FRET states that result in the observation of the most variance across different experimental conditions. However, it is difficult to compare smFRET molecules from different conditions without first binning into specific FRET states. Additionally, the

Nature Methods: doi:10.1038/nmeth.3602

dataset for the spliceosome was the first of its kind, leaving much unknown about the most important or probable FRET states within the data. We therefore chose the maximum number of reliable FRET states we can observe in practice given our resolution. However, once the most interesting or important FRET states are determined, the user can choose those or any FRET states for re-binning.

The clustering of single molecule trajectories will continue to improve as the HMM fitting tools for single molecule FRET data become more sophisticated. Incorporating multivariate HMM analysis that incorporates the donor and acceptor trajectories into the creation of HMMs may improve the results[2], although in many instances such as ours the analysis of the FRET signal directly was able to capture previously known biologically relevant conformations and reveal new ones. This allowed us to analyze a large quantity of trajectories since doing analysis on the FRET signal reduces the amount of data processed for each HMM calculation. Because the FRET signal is bounded between 0 and 1, whereas donor and acceptor intensities are not, this approach also allows us to more readily create bins of states that are easily compared. Furthermore, differences in fluorescence intensity due to slight variations in detector efficiency, illumination conditions, etc. can lead to different levels of fluorescence intensity across molecules that adopt the same FRET state.

Ideally, the TP matrix will perfectly describe the relative occupancies of the states of a given HMM. Unfortunately, the physical limitation of photobleaching shortens the observation window for each molecule and thus the amount of available information to create the HMM, making this assumption no longer valid. By combining the state occupancies with the TP matrix to create the FSM we create a matrix that is now weighted appropriately by the true occupancy in each state.

Nature Methods: doi:10.1038/nmeth.3602

Supplementary Note 2: SiMCAn Analysis of Early Splicing Complexes

We subjected the entire dataset of 8 experimental conditions to global analysis by SiMCAn. Application of SiMCAn revealed a disperse set of dynamics and cluster occupancies in the early splicing conditions ΔATP-WCE(WT) and ΔU6-WCE(WT) that stall at the CC2 and A complexes, respectively (**Supplementary Figs. 11 and 13**). Starting with a condition that favors formation of commitment complex 2 (CC2) by depletion of ATP (ΔATP-WCE(WT)), SiMCAn revealed clusters 0.75-**S** and 0.65-**S** of clade VI as the dominant clusters representing a conformational state that increases over our time course (**Supplementary Fig. 11**), indicating that the 5'SS and BP of the substrate are in close proximity. Such a behavior is expected for Ubc4 pre-mRNA, which contains a highly secondary structured intron with proximal 5'SS and BP[3]. Given that Ubc4 is able to efficiently form CC2 upon incubation with extract depleted of ATP (**Supplementary Fig. 15**), this also suggests that binding of U1 snRNP and BBP/Mud2 in CC2 is not sufficient to disrupt this secondary structure, which places the 5'SS and BP potentially close enough, but not properly positioned, for first-step catalysis. A group of dynamic clusters containing 0.75 and 0.65 as the most dominant states (clades III and V) was also significantly enriched (**Supplementary Fig. 13**), potentially signifying reversible binding and unbinding of the U1 snRNP and BBP/Mud2 to the pre-mRNA. However, such binding remains transient without the availability of ATP to activate the DExD/H-box ATPase Prp5 and load the U2 snRNA-protein complex (snRNP) onto the BP.

Accordingly, upon addition of extract containing ATP but depleted of U6 snRNA (ΔU6-WCE(WT)) to favor the A complex, SiMCAn identified a time-dependent increase a low-FRET 0.05-**S** cluster (clade I), indicating disruption of Ubc4's secondary structure and undocking of its 5'SS and BP (**Supplementary Figs. 11 and 13)**. This finding is consistent with the proposal that pre-mRNAs do

Nature Methods: doi:10.1038/nmeth.3602

not sample a proximal 5'SS-BP conformation until a later stage in spliceosome assembly after incorporation of the U5·U4/U6 tri-snRNP upon formation of the activated spliceosome (B^{act} complex)[1]. Notably, the preceding CC2 complex shows low occupancy in the 0.05-**S** cluster (**Supplementary Fig. 13**), further supporting the notion that its adoption requires an ATP-dependent assembly event. In this low-FRET state, the 5'SS and BP are stably undocked from one another, preventing premature catalysis prior to proper recognition and proofreading by the spliceosome. The dynamic clusters of clades III and V were again found to be moderately populated (**Supplementary Fig. 11**). Interestingly, several clusters in clade III appear to decrease over time as a result of the increase in occupancy of the 0.05-**S** cluster. This most likely indicates that reversible excursions are intrinsic to the complex, but can be biased towards a particular conformation upon activation of an ATPase. Furthermore, SiMCAn identified a significant population of molecules under A complex conditions that remained in the 0.75-**S** and 0.65-**S** clusters of clade II, characteristic of CC2 (**Supplementary Fig. 13**). It is likely that these molecules were not properly assembled into A complex and remain in a CC2-like state, consistent with the expected incomplete progression through the splicing cycle (**Supplementary Figs. 4 and 15**).

Supplementary Note 3: 3'SS-dependent differences prior to first step splicing

Comparison of the WT with the 3'SS substrate under ΔPrp2-WCE conditions revealed enrichment not only of the 0.05-**S** cluster, characteristic of molecules in an A-like conformation, but also of the 0.65-**S** and 0.75-**S** clusters specifically in the case of the 3'SS mutant (**Fig. 5**). The latter two clusters are characteristic of the CC2 complex and decrease over time. Previous work on other substrates has suggested that the identity of the 3'SS does not affect assembly of the spliceosome and that recognition and proofreading do not occur until after the first step of splicing[4]. Our results indicate that

Nature Methods: doi:10.1038/nmeth.3602

Ubc4 may behave slightly differently, perhaps due to its altered secondary structured relative to the WT[1]. Deletion of Prp2 may give the spliceosome ample time in the B^{act} stage to detect and discard or reverse-assemble on the 3'SS mutant substrate. Yet, once assembly is allowed to proceed unimpeded past the first step to the C complex, the spliceosome no longer has sufficient time to detect and discard the mutant substrate or reverse-assemble on the substrate. As a result, the subtle differences in assembly become muted and the two substrates behave more similarly (**Fig. 5**).

Supplementary Note 4: SiMCAn-directed experimentation leads to discovery of a 3'SS-induced discard pathway after the first step of splicing

The initial set of experimentation used for input into SiMCAn involved smFRET data collected solely from incubation either the 3'SS or WT substrate with unmodified extract (WT-WCE(3'SS) and WT-WCE(WT), **Supplementary Fig. 18**). Clustering by SiMCAn resulted in formation of 5 dynamic clusters in addition to the 10 static clusters. Interestingly, cluster 0.05-**S** became increasingly enriched with the 3'SS mutant but remained nearly constant with the WT substrate. This initial set of exploratory research led us to a hypothesis that the 3'SS substrate adopts a static, low FRET conformation after the Bact complex that we wanted to further test. We thus later added in the Prp16DN experiments to determine if this low FRET conformation is formed before or after the first step of splicing.

With the now complete dataset, SiMCAn again identified a 0.05-**S** cluster (clade I) as particularly enriched in the 3'SS mutant substrate after the Prp16-dependent reorganization of the spliceosome. The ATPase Prp16 is known to crosslink to the 3'SS and is required for formation of a functional step 2 active site immediately following the first step of splicing[5]. In addition, one of the

Nature Methods: doi:10.1038/nmeth.3602

second-step splicing factors, Slu7, a protein known to also bind mutant 3'SS, was proposed to be involved with efficient docking of the 3'SS into the step 2 active site[6]. The deficiency in docking observed with the 3'SS may be the result of Slu7 and other second-step factors preventing docking, or the result of the ATPase activity of Prp22. In this latter case, the 3'SS may transiently dock into the second step conformation, but Prp22 rapidly recognizes and discards the mutated 3'SS. Either hypothesis would constitute a form of proofreading and explain the accumulation of a discarded, undocked substrate unable to proceed through the second step of splicing. These findings also provide further evidence that SiMCAn is an excellent form of exploratory analysis capable of generating experimentally testable hypotheses. Additionally, we have evidence that this low-FRET conformation to be substrate that is no longer bound by spliceosomal components. If the spliceosome were to disassemble upon completing the first step of splicing and encountering the mutated 3'SS, the free 5' exon would be lost so that the pre-mRNA molecule would not show a Cy5 signal and thus would not have been selected for analysis as an RNA molecules containing both fluorophores. Furthermore, incubation of the fluorophore-labeled Ubc4 substrate with splicing buffer alone and no spliceosomal components reveals a dominant high-FRET peak with a smaller low-FRET population featuring very little zero FRET (**Supplementary Fig. 19**). Therefore, the zero-FRET state adopted after the first step of splicing with the 3'SS substrate is most likely still bound by the spliceosome.

Supplementary Note 5: The mathematical underpinnings of SiMCAn are based off standard techniques

The fitting of the smFRET traces to HMMs directly uses the vbFRET algorithm[7]. The values for the best fitting HMM assigned for every trace is then rescaled to the closest of the ten evenly spaced FRET values spanning the range of 0.5 to 0.95 with increments of 0.10. These rescaled traces are

Nature Methods: doi:10.1038/nmeth.3602

then used to construct the transition probability matrices in the standard way of counting the number of transitions between states and time points without transitions. The transitions from transition probability matrices are scaled to unity. The percent occupancy of the fret states describing the trace is appended to the transition probability matrix to create the FRET similarity matrix. The FRET similarity matrices are then clustered together using Ward's method for hierarchical clustering where the distance measurement between clusters *r* and *s* is given by:

$$d(r,s) = \sqrt{\frac{2n_r n_s}{n_r+n_s}}\left(\bar{x_r^2} - \bar{x_s^2}\right)^{\frac{1}{2}}, \qquad \text{Eq. 1}$$

where n is the number of elements and x is the centroids for the respective clusters *r* and *s*. This clustering algorithm is natively implemented using Mathwork's MATLAB software. The cutoff for the number of clusters is determined by k-means described by:

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

where B is the number of randomly cluster sets and W^r_k is the average inter-cluster distance in a random set of k clusters while W_k is the average inter-cluster distance of the actual clusters. The resulting clusters are then characterized by the average transition probability matrix and probability distribution of FRET states.

References

1. 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).
2. Liu, Y. *et al.* A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis. *J. Phys. Chem. B* **114**, 5386-5403 (2010).
3. Abelson, J. *et al.* Conformational dynamics of single pre-mRNA molecules during in vitro splicing. *Nat. Struct. Mol. Biol.* **17**, 504-512 (2010).
4. Schwer, B. & Guthrie, C. A conformational rearrangement in the spliceosome is dependent on PRP16 and ATP hydrolysis. *EMBO J.* **11**, 5033-5039 (1992).

Nature Methods: doi:10.1038/nmeth.3602

5. 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).
6. 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).
7. 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).

Nature Methods: doi:10.1038/nmeth.3602