Decrypting the Information Exchange Pathways across the Spliceosome Machinery

Andrea Saltalamacchia,[1] Lorenzo Casalino,[2] Jure Borišek,[3] Victor S. Batista,[4] Ivan Rivalta,[5,6] and Alessandra Magistrato[7]*

1 International School for Advanced Studies (SISSA/ISAS), via Bonomea 265, 34136, Trieste, Italy

2 Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, U.S.A.

3 National Institute of Chemistry, Hajdrihova 19, SI-1001, Ljubljana, Slovenia

4 Department of Chemistry, Yale University, New Haven, CT 06520, U.S.A.

5 University of Bologna**,** Dipartimento di Chimica Industriale "Toso Montanari", Viale del Risorgimento 4, Bologna, Italy

6 Univ Lyon, Ens de Lyon, CNRS UMR 5182, Université Claude Bernard Lyon 1, Laboratoire de Chimie, F69342, Lyon, France

7 Consiglio Nazionale delle Ricerche–Istituto Officina dei Materiali, International School for Advanced Studies (SISSA), via Bonomea 265, 34135, Trieste, Italy

3. SUPPORTING TABLES

4. REFERENCES

1. SUPPLEMENTARY RESULTS

In order to identify the critical protein regions responsible of the C complex functional dynamics, we initially computed the cross-correlation matrix (CCM) based on Pearson's correlation coefficient (CCs) from the combined replicas trajectories (Figure S1). We have also verified that the results obtained from the combined replica trajectory were qualitatively similar to the single replicas (Figure S2). [1–5] For clarity reasons, we simplified this rough CCM into a coarse-grained matrix (Figure S1). This has been done by summing the CCs of each pair of protein/domain and by averaging it by the corresponding number of residues, obtaining pairwise correlation scores (CSs). The resulting coarse-CCM shows that Prp8 is divided in two dynamical regions: (i) one comprises the endonuclease and RNAse-H domains, which negatively correlate with the rest of Prp8, with all RNA filaments, and all the rest of the system, but Clf1, Cwc25 and Ecm2. (ii) The second region encloses all other Prp8 domains, which positively correlate with all RNA filaments.[4] Of note Clf1, a protein composed entirely of extended Half-A-Tetratrico Peptide (HAT) Repeats and suggested to promote the spliceosome assembly,[6] protrudes from the N-term of Prp8 towards Cwc2 and over Prp46 and strongly correlates in lock-step motion with the RNAse-H/endo domains, even though these being separated by a huge distance. This suggests that information exchange is taking place between them and that the repetitive structure of Clf1 may be instrumental for signal transfer during splicing.[5,7] Due to the prolonged and repetitive architecture of Clf1, we plotted the coarse CCM by splitting its structure into its constituent HAT repeats. This allowed us to pinpoint a switch from positive to negative correlation in the CCM between HAT-repeat 2 and 3 (H2-3), which most probably indicates the presence of a hidden hinge, modulating the Clf1's functional dynamics. While the CCM based on Pearson coefficients allows to quickly pinpoint positively and negatively correlated motions, it also lacks a fraction of correlation due to non-linear and non-parallel motions. Therefore to dissect more accurately the critical mode of information exchange we also exploited the mutual information approach,[8] a more effective method proven to capture all types of correlations, after verifying the Pearson-based CCs and the linear mutual information correlation coefficients ($^{\mathrm{LMI}}$CCs) have a close correspondence between the most highly correlated regions (Figure S5). The linearized version of the mutual information has been used for computational reasons (see Methods).

2. SUPPORTING FIGURES

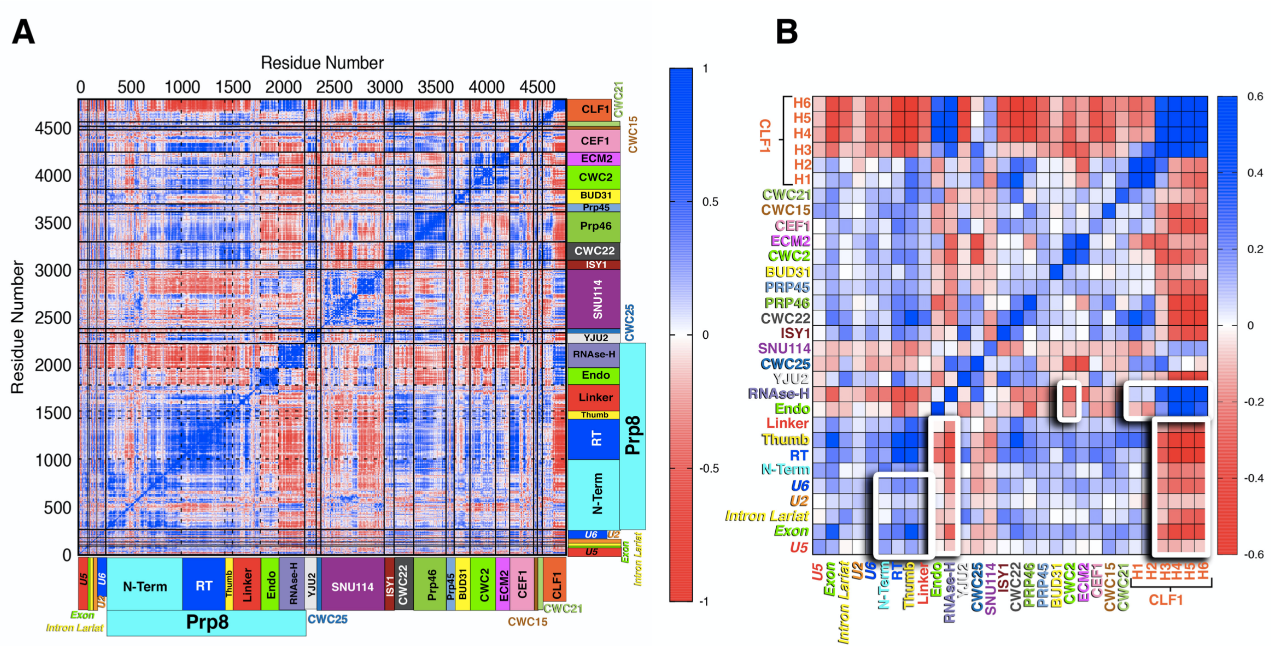

Figure S1. (A) Cross-correlation matrix based on per-residue Pearson's correlation coefficients (CCs) as derived from the mass-weighted covariance matrix calculated over 3-replicas of classical molecular dynamics trajectories. CCs values range from -1 (red, anti-correlated motions) to +1 (blue, correlated motions). (B) Coarse matrix summing the correlation scores and normalizing over the products of the residues of the considered SPL proteins/domains. Pairwise correlation scores (CSs) are reported in the range from -0.6 to 0.6 for clarity reasons. In green are encircled regions relevant to explain the functional movements captured by the principal component analysis. Protein names and their domains are labeled with the same color code of Figure 1.

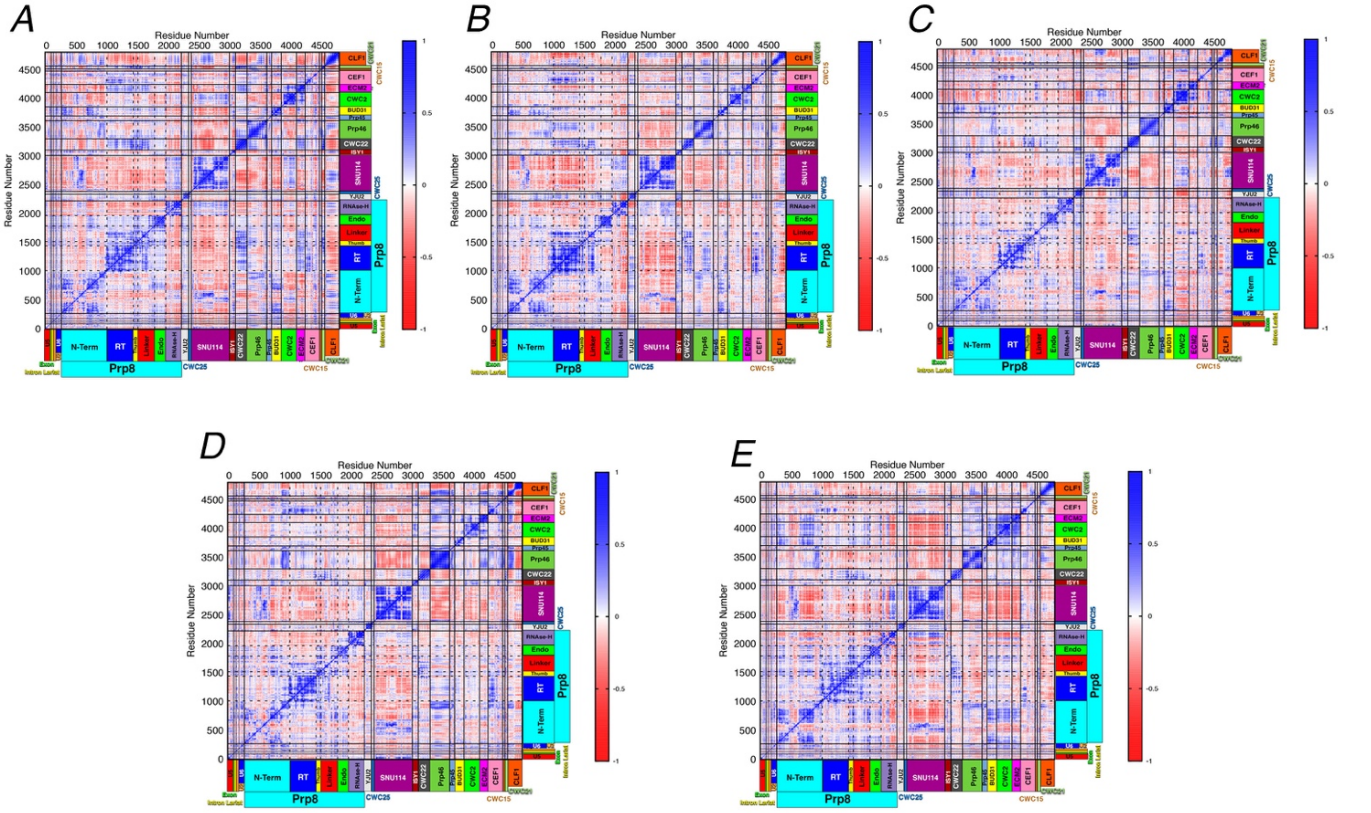

Figure S2. Per-residue Pearson's cross-correlation coefficients (CCs) derived from the mass-weighted covariance matrix calculated over the last 700 ns of MD trajectories for the 5 replicas of the C model in (A)-(E), respectively. CCs values range from -1 (read, anti-correlated motions) to +1 (blue, correlated motions). The protein names and their domains are reported on the bottom and on the left side, highlighted with boxes of different colors.

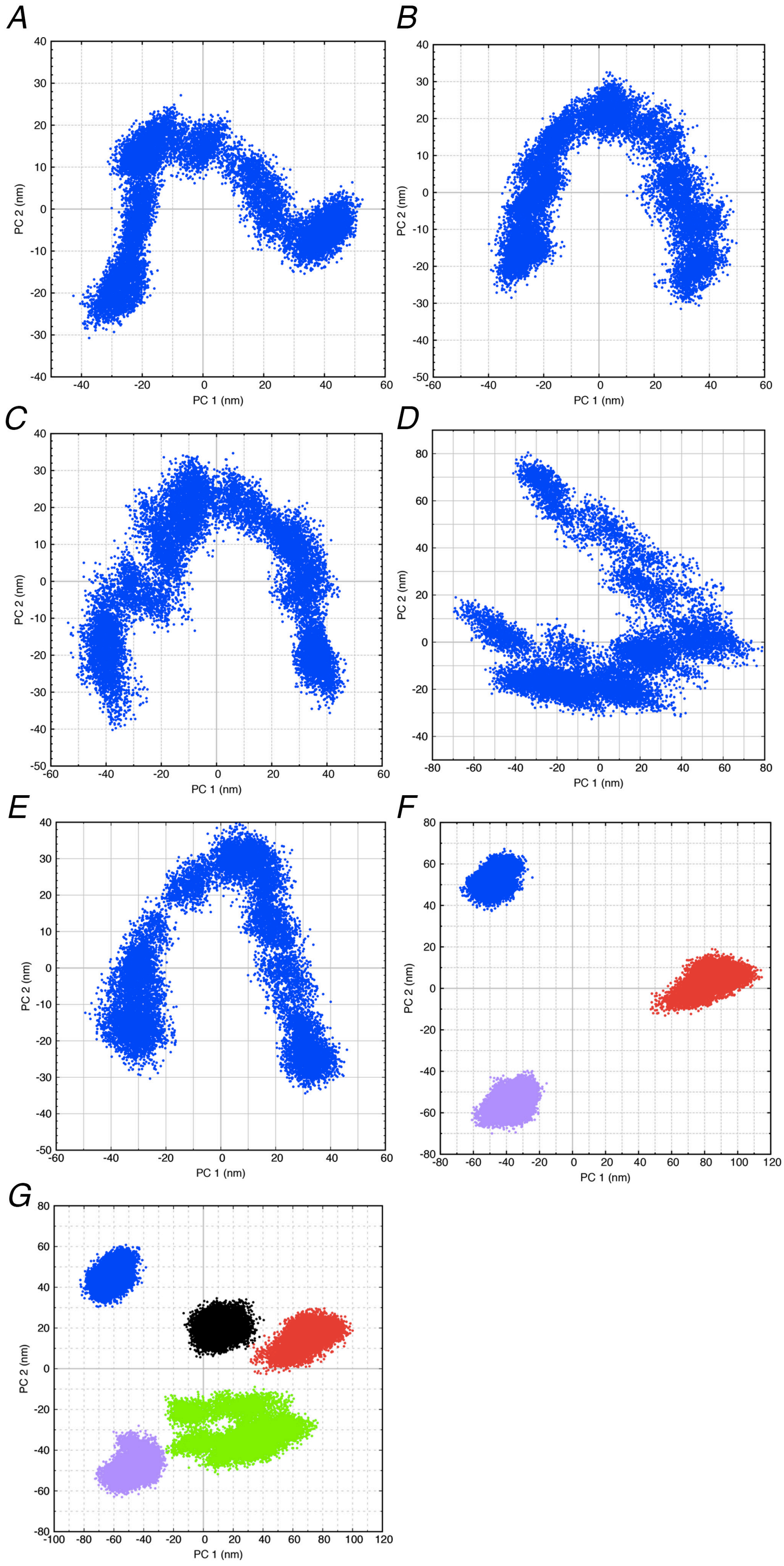

Figure S3. Scatter plot of Conformational subspace of PC1 vs PC2 for replica 1 to 5 in (A-E), respectively, merged replicas 1-3 (F), and merged replicas 1-5 (G).

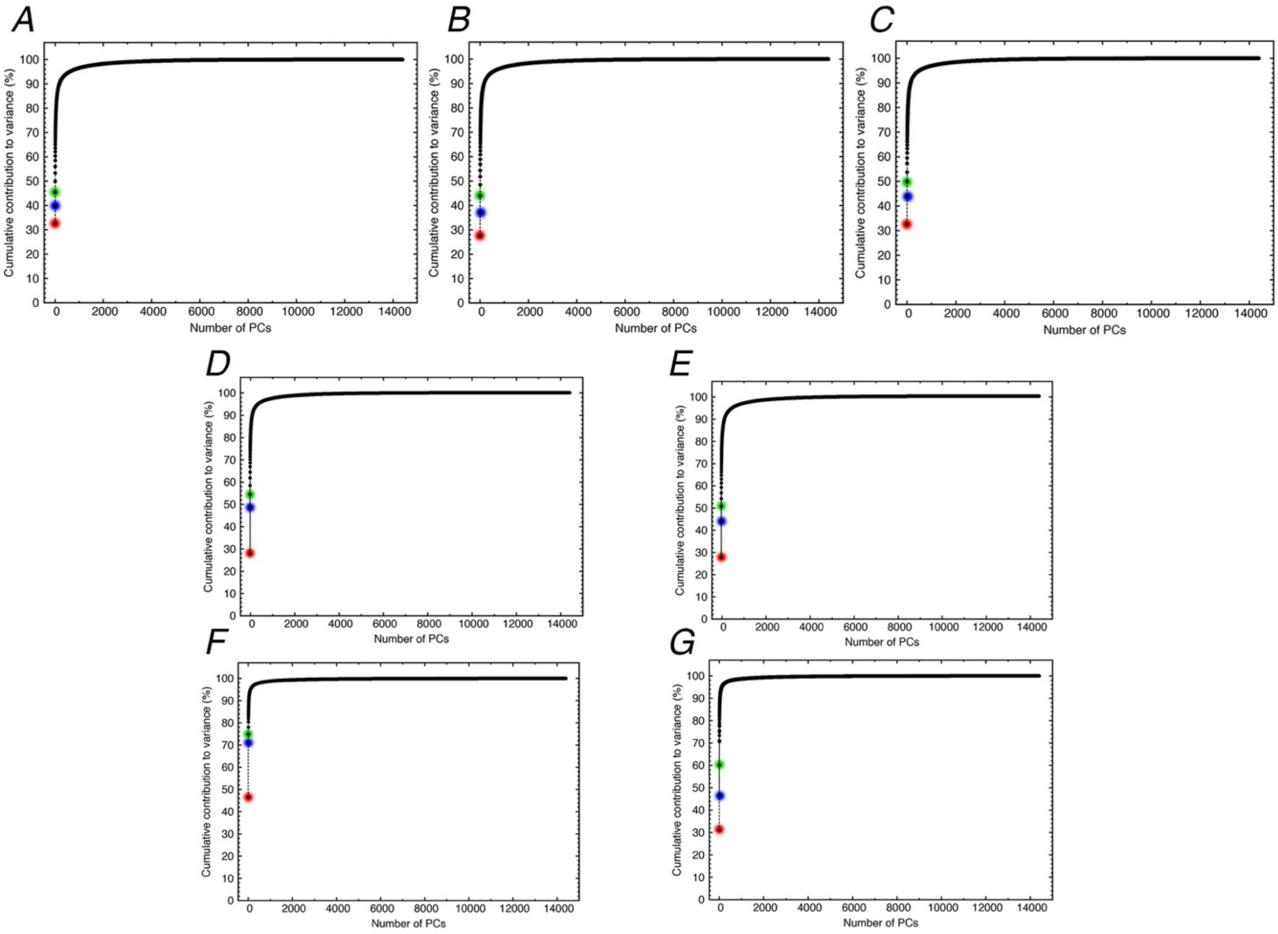

Figure S4. Principal components (PCs) cumulative contribution to variance for (A) replica 1, (B) replica 2, (C) replica 3 and (D) replica 4, (E) replica 5, (F) 3 replica combined trajectory, (G) 5 replicas combined trajectory. On y-axis is depicted Cumulative contribution of PCs (x-axis) to the variance of the overall motion calculated upon Principal Component Analysis. The contributions from the first three PCs are highlighted in red, blue and green, respectively.

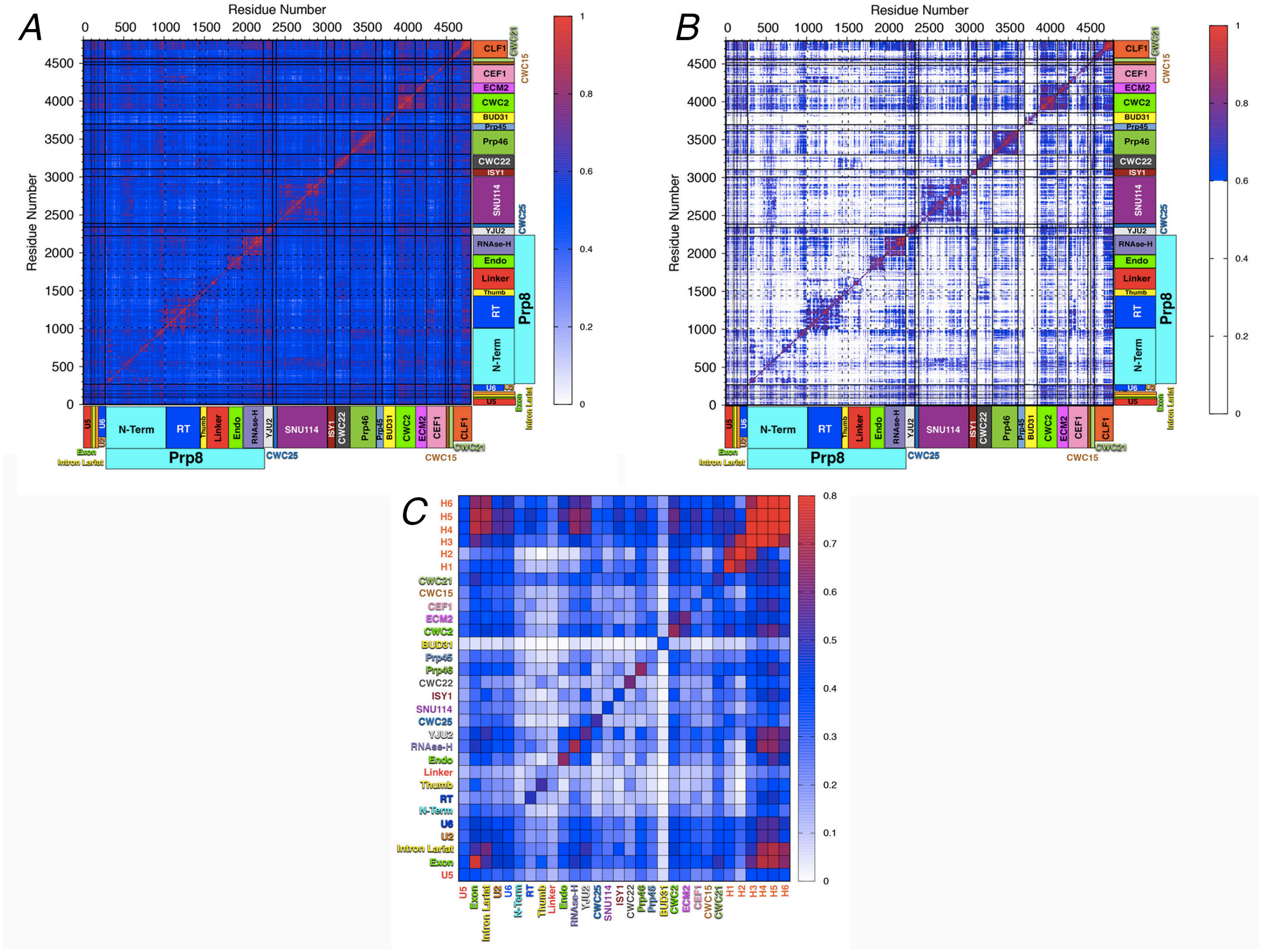

Figure S5. (A) Linear Mutual Information correlation coefficients (^{LMI}CCs) for the combined 3 replicas trajectory. These values range from 0 (white, no correlated motions) to +1 (red, correlated motions). (B) Linear Mutual Information matrix after filtering the correlation coefficients below 0.6. (C) Coarse grained matrix of pairwise correlation scores (^{LMI}CSs) given by summing ^{LMI}CCs of each pair of protein/domain and averaging by the corresponding number of residues, after filtering all values below 0.6. ^{LMI}CSs are reported in the range from 0 to 0.8 for clarity reasons. Protein names and their domains are labeled with the same color code of Figure 1 of the main text. Multiple strong correlations are clearly visible in this matrix confirming the pivotal role of Prp8 in establishing the intricate correlation network among its domains, which direct the SPL motion. As well, Clf1 shares many strong correlations, in particular, with the RNAse-H, Endo domains of Prp8 and Cwc2.

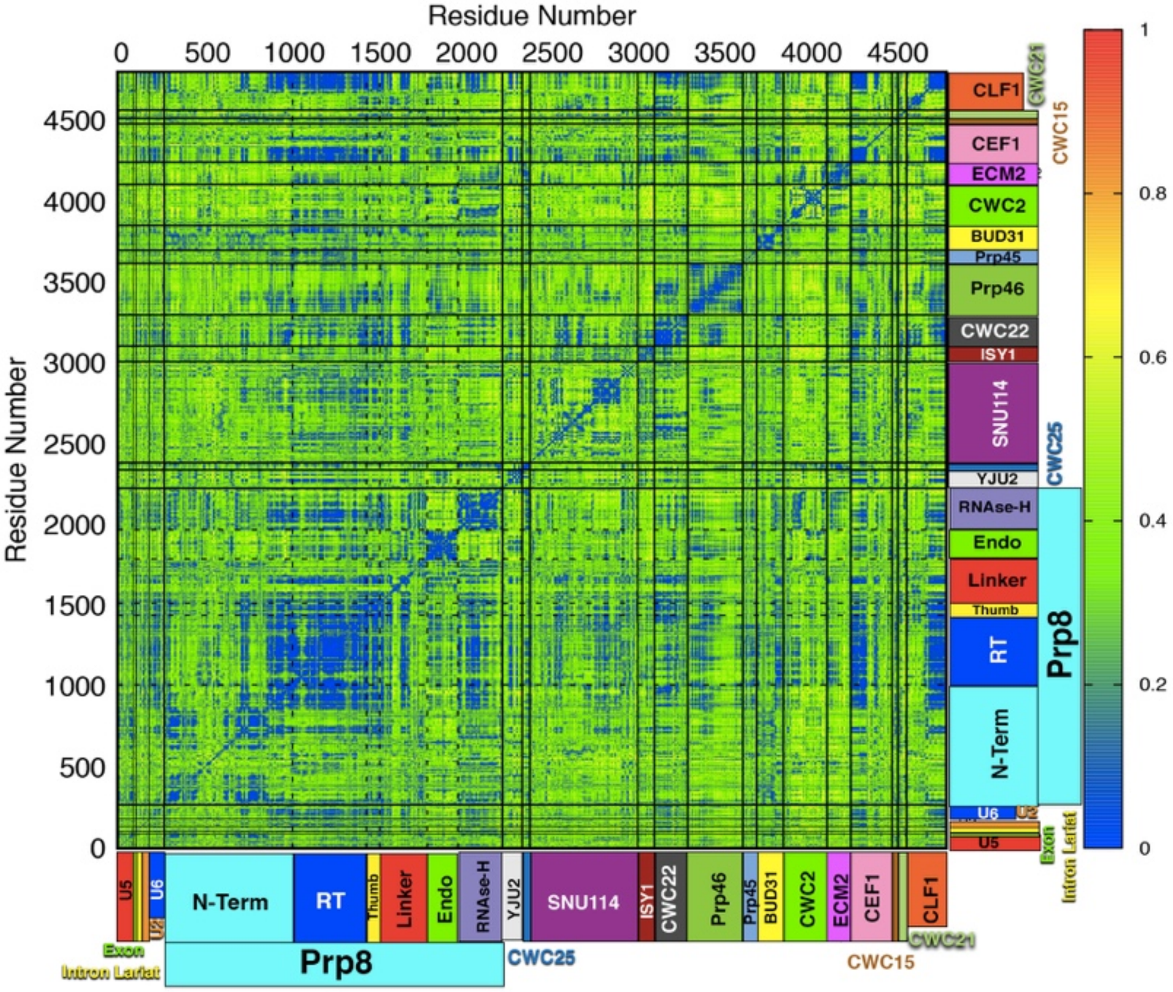

Figure S6. Root Mean Square Deviation matrix between ^{LMI}CCs and Pearson CCs for the combined 3 replicas trajectory. These values range from 0 (blue, low RMSD i.e. same values) to +1 (red, high RMSD i.e. completely different values). Protein names and their domains are labeled with the usual color code of Figure 1 of the Main Text.

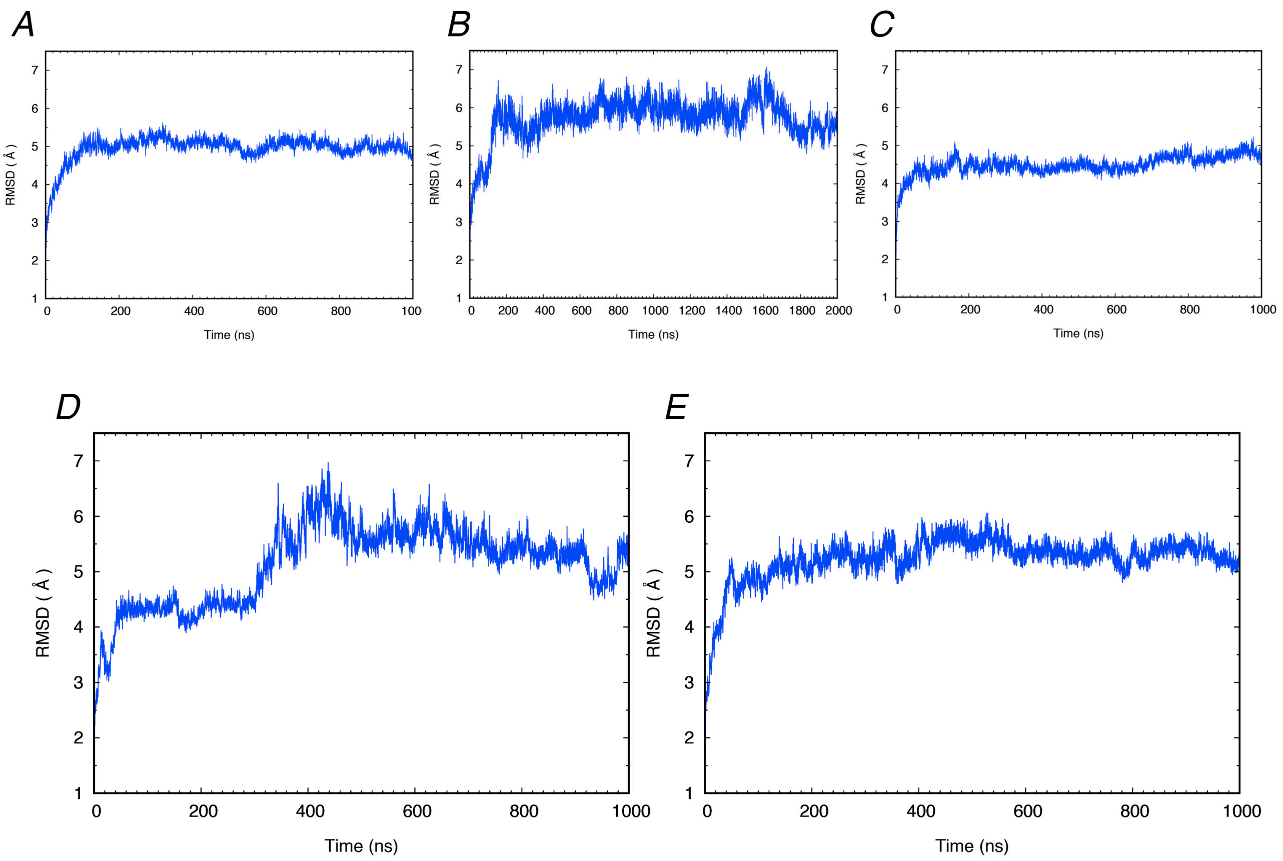

Figure S7. Root Mean Square Deviations (RMSD) vs. simulation time (ns) calculated on the production phase of molecular dynamics trajectories for the four 1-μs-long replicas of C model and for replica 2 prolonged up to 2-μs.

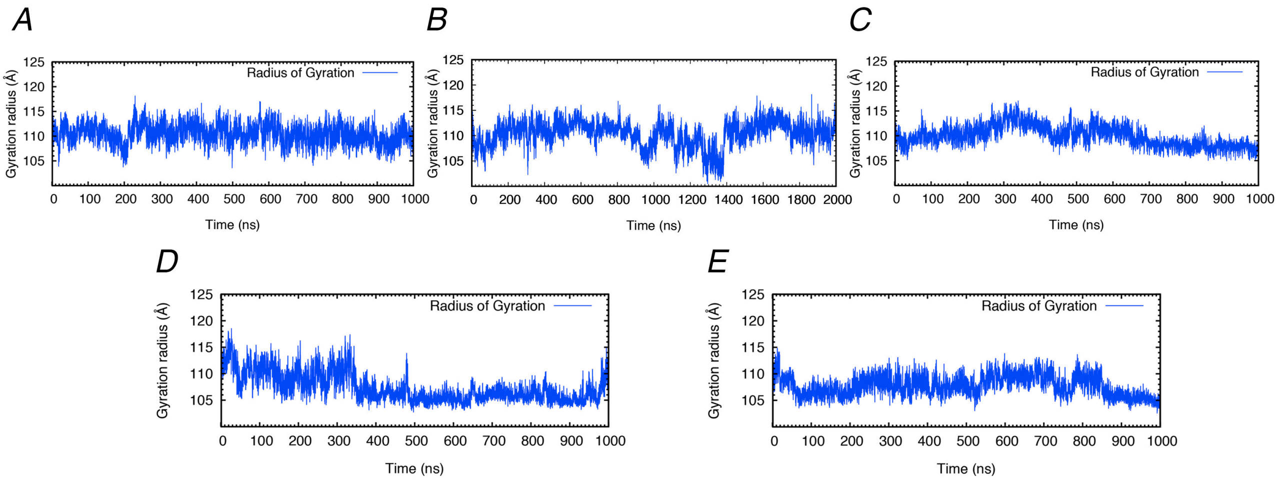

Figure S8. Radius of gyration vs. simulation time (ns) calculated on the production phase of molecular dynamics trajectories for the four 1-μs-long replicas of C model and of replica 2 prolonged up to 2-μs.

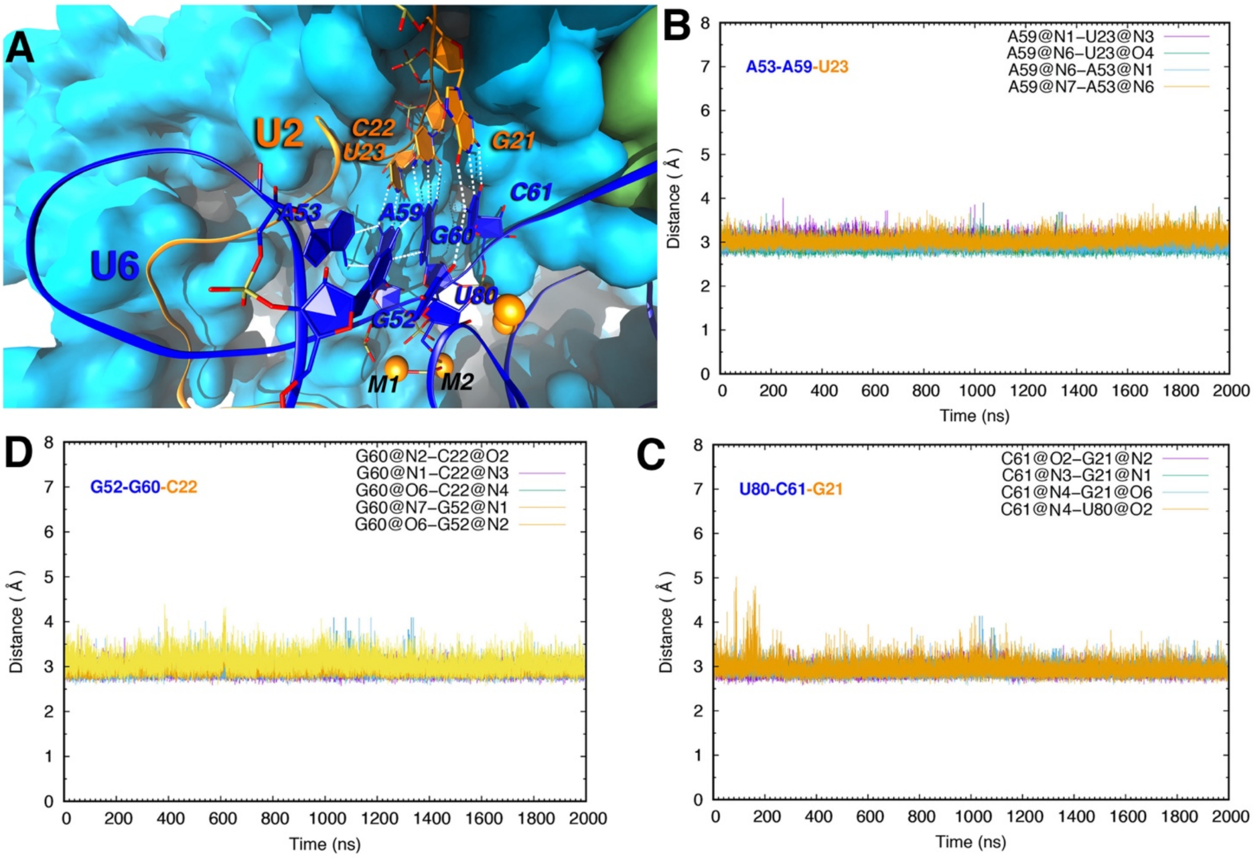

Figure S9. (A) Representative snapshot of the catalytic site grafting the triple-helix made by U2/U6 snRNAs. U2 and U6 snRNAs are represented as orange and blue tubes, respectively. Mg^{2+} ions are depicted as orange spheres with the catalytic Mg^{2+} ion labeled as M2. The nucleotides involved in the triple-helix are pictured in licorice. The Prp8 protein is shown in cyan surface. Hydrogen bonds (H-bonds) between RNA base-pairs are depicted as white dash lines. (B) Time evolution (ns) of the H-bonds distances (Å) between base-pairs of the nucleotides involved in the triple-helix in the longest replica, showing that the structural integrity is maintained during all simulation time. In all simulations, the triple helix architecture of the active site remained well preserved (Figure S3), with the 5 Mg^{2+} ions engaging strong interactions with the phosphate groups of U6 snRNA and being nested within a positively charged pocket formed by Prp8.

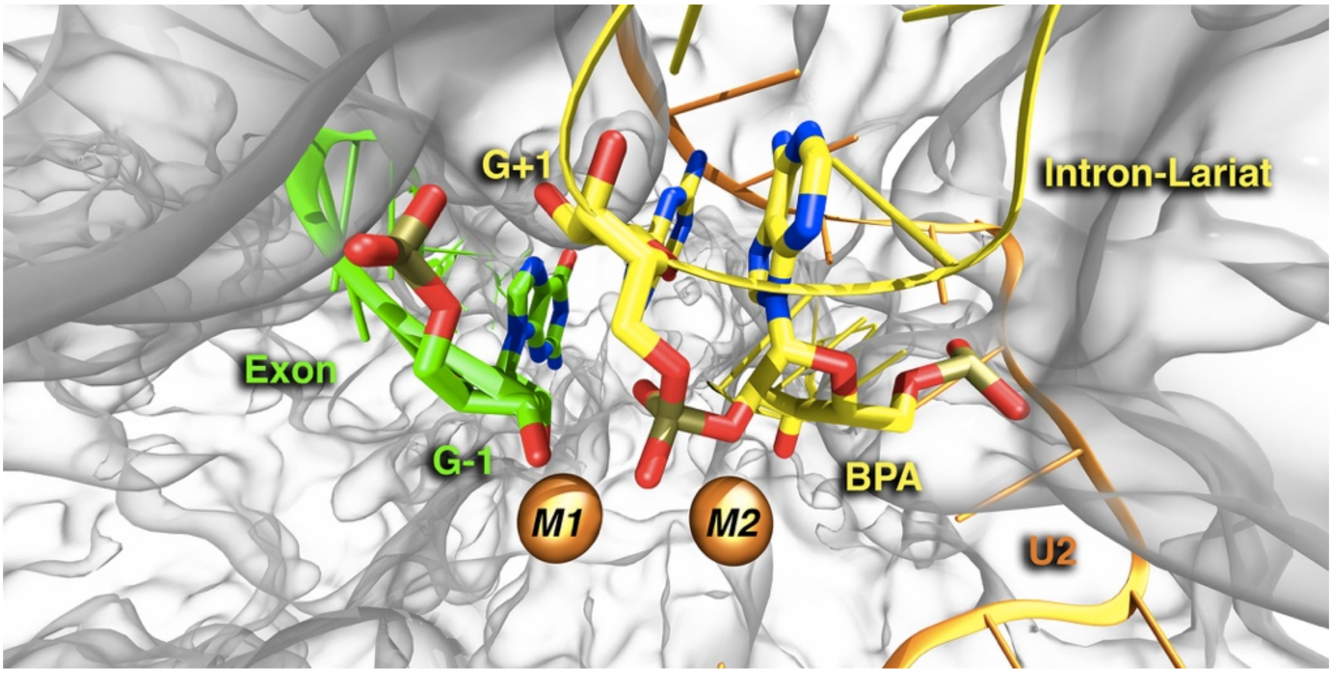

Figure S10. 5'Splicing-Site showing the truncated 5'exon in green and the Branching Point Adenosine bound to the first intronic base (G+1) via the non-canonical 5'phosphate-O2'oxygen bond. The two Mg^{2+} ions are in orange van der Waals spheres. During the MD simulations, the O3' extremity of the cleaved 5'-exon remains at an average distance of 3.2 Å from the 5'-phosphate of the intron nucleotide G(+1).

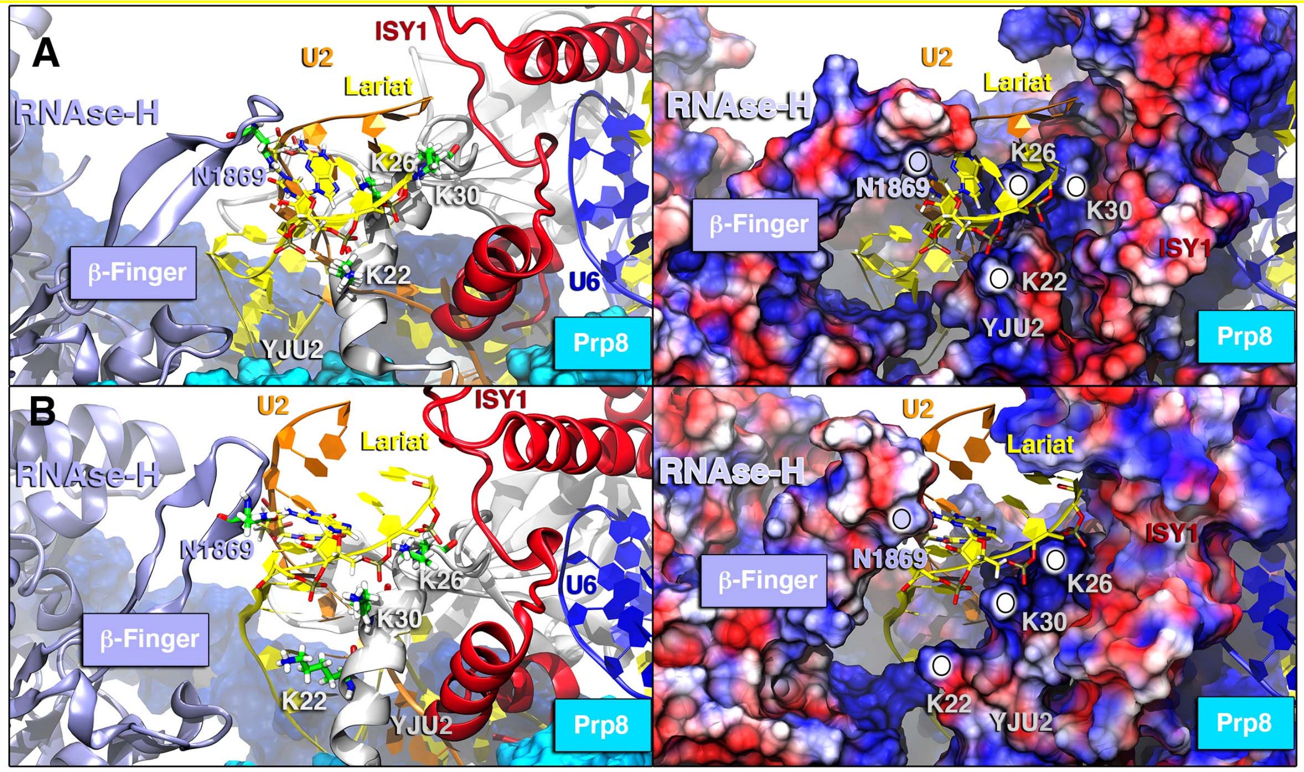

Figure S11. Structural rearrangement of the U2/intron lariat helix during PC1 (A and B), and electrostatic potential of representative frames extracted from the essential dynamics trajectory underlying the RNAse-H movement and the consequent IL/U2 helix wrapping. This is mediated by the Asn1869 of the ß-finger and by the Lys22, 26 and 30 of Yju2.

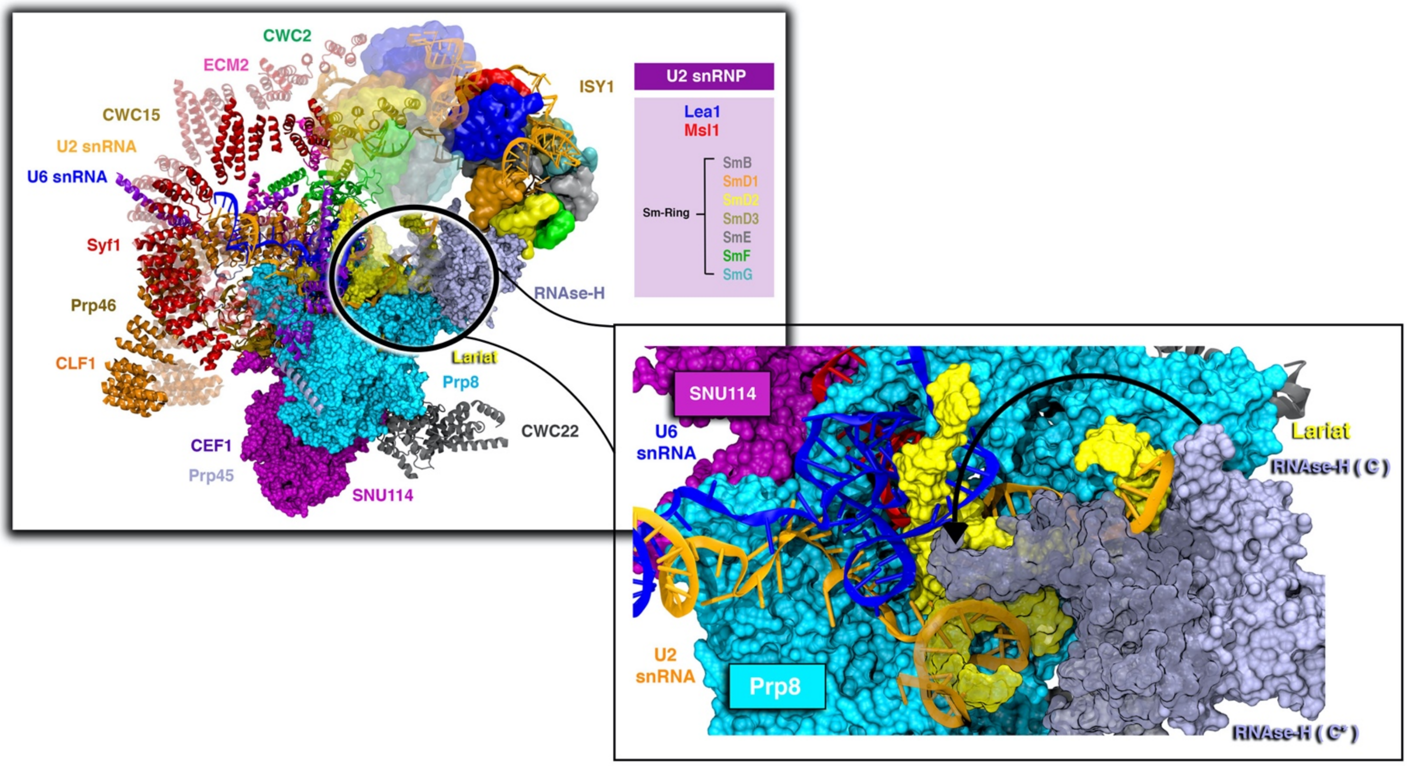

Figure S12. Experimental structures alignment of C and C* complex focusing on the components affected by conformational changes from one intermediate to the the other. In transparency are represented Syf1, Clf1 and U2snRNP of C* components. Darker objects are representing C proteins. CWC2 in green, Clf1 in orange (on the left), Syf1 in red and U2snRNP in plastic-surface, lariat in yellow and RNAseH in iceblue (on the right). U2snRNP Sm-Ring gets away from the RNAse-H domain by a rotation of Clf1 and Syf1. The insect shows the C* conformation of lariat and RNAse-H represented in transparent darker Surf, where the ß-finger of the RNAseH domain embraces the intron/U2 helix, interacting with its minor groove.

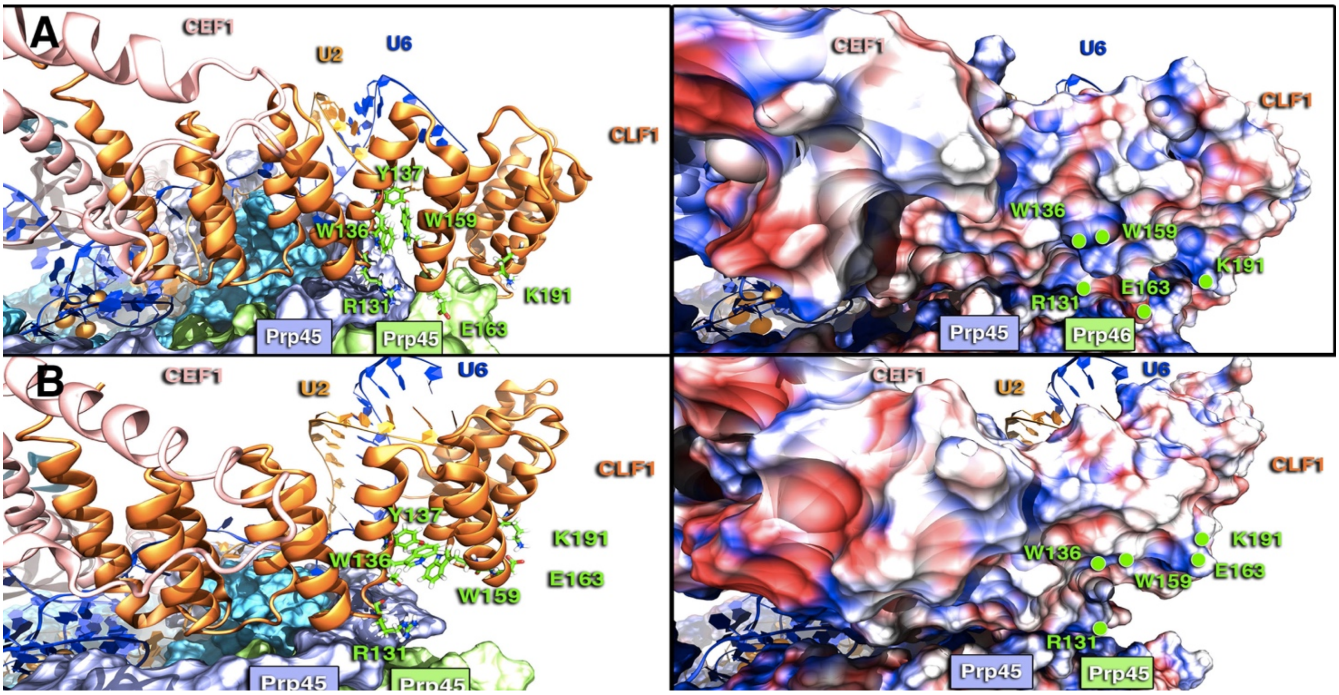

Figure S13. The electrostatic hinge of Clf1 and the rearrangements of hydrophobic interactions. (Left)Prp45 (light green), Prp46 (lilac) and Prp8 (cyan) are represented with surface; U2 (orange) and U6 (blue) are shown as New Ribbon. (Right) Proteins are shown with electrostatic surface (blue/red colors for positive/negative charges, respectively). The motions of Clf1 appears to be modulated by a rearrangement of salt-bridge interactions between E163 and K191, while the plasticity of the hinge located at H2-3 is associated to a reorganization of extended π-stacking interactions involving Y137, W159, W136. (A) Namely, Y137 establishes a T-shaped stacking with W159, which π-stacks with W136 through a parallel-displaced conformation. (B) After the functional rearrangement, W136 forms a T-shape stacking with Y137, inducing a downstream displacement of the nearby HAT-repeats.

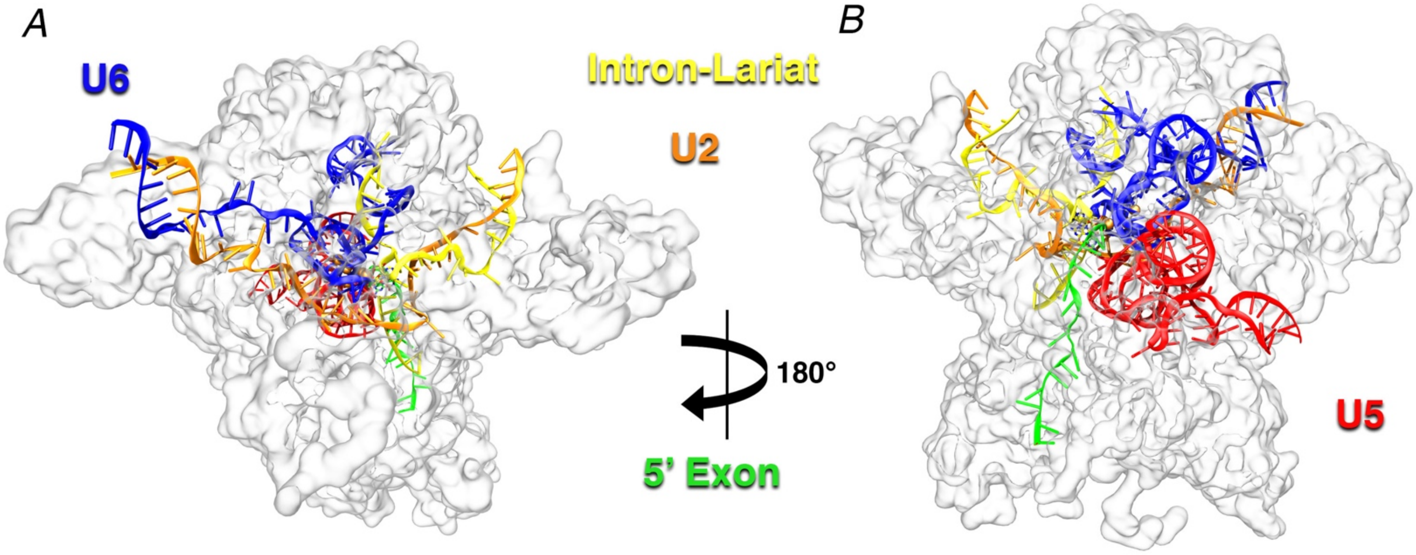

Figure S14. Small nuclear (sn)RNAs elements within the spliceosome. Front (A) and back (B) views of RNAs positions. In transparent surface is depicted the protein counterpart of SPL and 5'Exon, Intron-Lariat, U6 snRNA, U2 snRNA, U5 snRNA are displayed in green, yellow, blue, orange and red, respectively.

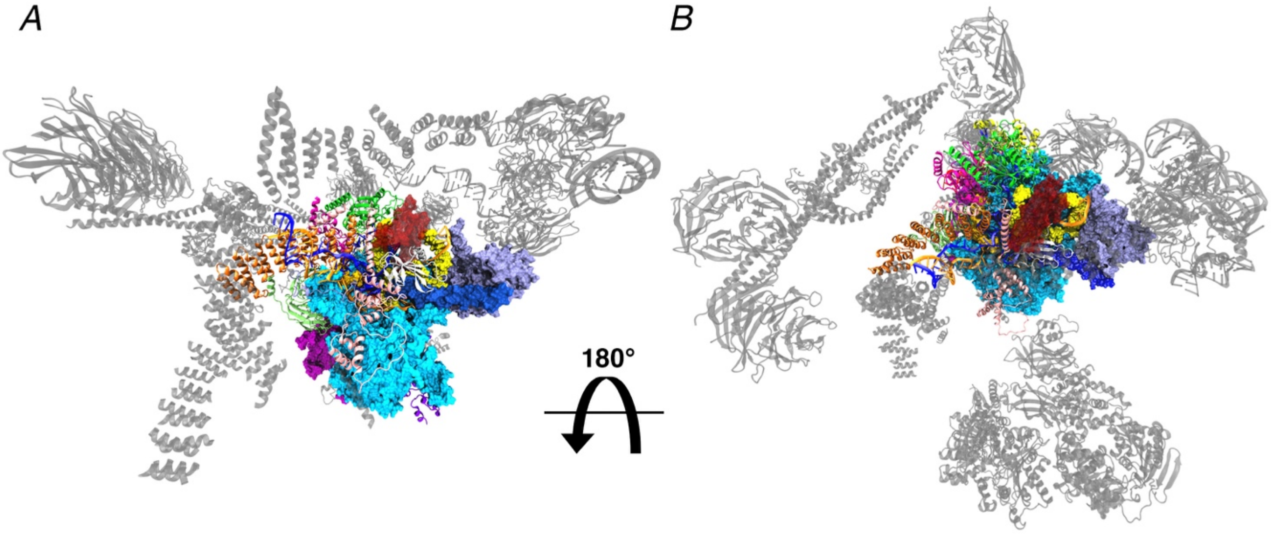

Figure S15. Overall structure of C complex (PDB ID: 5LJ5[9]) front (A) and top (B). In grey are displayed all the proteins that were not included into the model for the MD simulation.

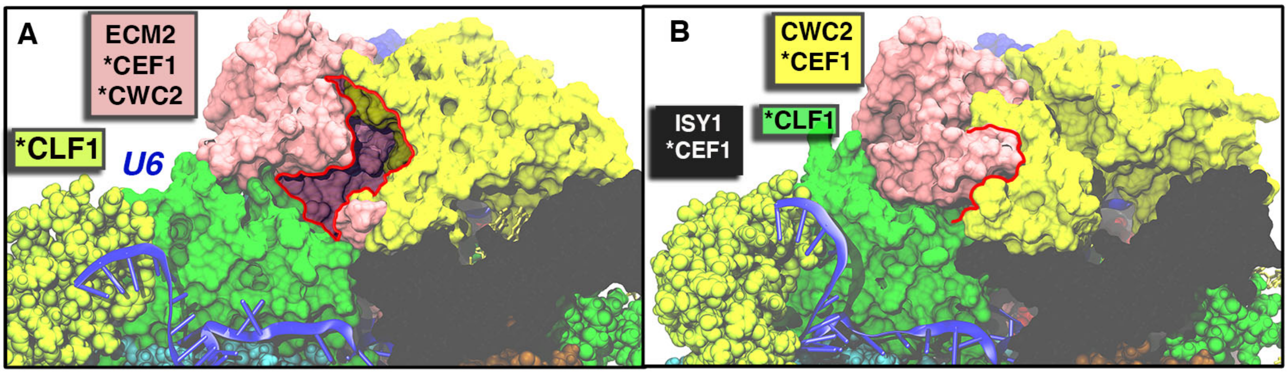

Figure S16. Possible binding pocket lying on the communication path II and its open (A) or closed (B) states of the 'hammer-like' motion described by PC1. Small molecules targeting this region could critically interfere with the internal communication network underlying the spliceosome dynamics. Asterisks indicate domains or proteins spread in distinct communities. Proteins are represented with the same color code of Figure 3C.

3. SUPPORTING TABLES

Table S1. Protein/RNA composition of the SPL. CONSIDERED column stand for the residues included in the simulation. MODELLED refers to residues modelled using MODELLER 9v16 [10]. RESOLUTION indicates the cryo-EM resolution for each protein/RNA

SPL C Complex Model				
PROTEIN NAME	CONSIDERED	MODELLED	RESOLUTION	LENGTH
U5 snRNA	28-53 + 62-125	/	3.8 – 7.6	90
EXON	-16 - -1	/	3.4 – 6.4	16
LARIAT INTRON	1-10 + 54-76	/	3.4 – 7.2	32
U2 snRNA	3-47	/	3.8 - 6.0	45
U6 snRNA	16-102	/	3.6 – 6.4	87
Prp8	128-2085	429-457	3.4 – 5.8	1958
YJU2 (CWC16)	2-115	/	3.8 – 5.4	114
CWC25	3-48	/	3.8 – 7.0	46
SNU114	71-693	516-533	3.8 – 7.2	623
ISY1	1-97	/	3.8 – 6.2	97
CWC22	289-481	400-413	4.6 – 8.2	193
Prp46	111-428	/	3.4 – 6.6	318
Prp45	104-184	151-158	4 - 8.4	81
BUD31	2-156	/	3.6 – 6.8	155

CWC2	3-254	/	3.6 – 6.0	252
ECM2 (SLT11)	6-144	93-100	4.0 – 7.0	139
CEF1	12-247	101-147	3.8 – 6.2	236
CWC15	7-42	/	3.6 – 7.6	36
CWC21	2-50	/	3.8 – 7.4	49
CLF1	37-273	/	3.8 – 6.4	237
Mg+	#5	Saxena Force Field		
Zn(2+)	#7	Pang Force Field		
NA+	#201	Joung & Cheatham FF		
Wat mol	#229850	TIP3P		
Total number of atoms (System) : 772679		**Protein Force Field:** ff12SB	**Cryo-EM:** 3.8 Å (5lj3)	
Solute atoms (SPL): 83129		**RNA Force Field:** ff99+bsc0+χOL3FF	**Organism:** *Schizosaccharomyces Cerevisiae*	

Table S2. List of residues lying along the communication pathways I and II (node betweenness >0.6). In bold are reported the key residues with node betweenness >0.85. Table cells are colored with the same color code of Figure 1 of the main text.

PATH I		PATH II	
Protein	Residue	Protein	Residue
CLF1	ARG62	CLF1	GLU89
CWC15	**SER13**	CEF1	ASP216
CWC15	ALA11	CEF1	TYR213
Prp8 (N-Ter)	LEU783	ECM2 (SLT11)	GLU103
Prp8 (N-Ter)	GLU788	CWC2	ALA118
Prp8 (N-Ter)	**CYS792**	CWC2	LYS116
Prp8 (N-Ter)	ALA795	CWC2	LEU109
Prp8 (RT)	MET1095	CWC2	ARG63
Prp8 (RT)	HIS1097	BUD31	PHE142
Prp8 (RT)	**ASN1099**	Prp8 (N-Ter)	**GLN558**
YJU2 (CWC16)	PHE97	Prp8 (N-Ter)	LEU192

YJU2 (CWC16)	ARG83	Prp8 (N-Ter)	**ASN203**
YJU2 (CWC16)	ILE81	Prp8 (N-Ter)	**THR205**
YJU2 (CWC16)	ILE79	Prp8 (N-Ter)	**ARG207**
CWC25	THR26	Prp8 (N-Ter)	**ILE209**
CWC25	LEU30	Prp8 (N-Ter)	**LEU318**
		Prp8 (N-Ter)	ASP651
		Prp8 (N-Ter)	ARG236
		Prp8 (Endo)	PHE1756
		Prp8 (Endo)	VAL1662
		Prp8 Endonuclease	ASP1664
		Prp8 (RNAse-H)	LYS1912

4. REFERENCES

(1) Palermo, G.; Miao, Y.; Walker, R. C.; Jinek, M.; McCammon, J. A. Striking Plasticity of CRISPR-Cas9 and Key Role of Non-Target DNA, as Revealed by Molecular Simulations. *ACS Cent. Sci.* **2016**, *2* (10), 756–763. https://doi.org/10.1021/acscentsci.6b00218.

(2) Pavlin, M.; Spinello, A.; Pennati, M.; Zaffaroni, N.; Gobbi, S.; Bisi, A.; Colombo, G.; Magistrato, A. A Computational Assay of Estrogen Receptor α Antagonists Reveals the Key Common Structural Traits of Drugs Effectively Fighting Refractory Breast Cancers. *Sci. Rep.* **2018**, *8* (1). https://doi.org/10.1038/s41598-017-17364-4.

(3) Ricci, C. G.; Silveira, R. L.; Rivalta, I.; Batista, V. S.; Skaf, M. S. Allosteric Pathways in the PPARγ 3-RXRα Nuclear Receptor Complex. *Sci. Rep.* **2016**, *6*. https://doi.org/10.1038/srep19940.

(4) Casalino, L.; Palermo, G.; Spinello, A.; Rothlisberger, U.; Magistrato, A. All-Atom Simulations Disentangle the Functional Dynamics Underlying Gene Maturation in the Intron Lariat Spliceosome. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (26), 6584–6589. https://doi.org/10.1073/pnas.1802963115.

(5) Borišek, J.; Saltalamacchia, A.; Gallì, A.; Palermo, G.; Molteni, E.; Malcovati, L.; Magistrato, A. Disclosing the Impact of Carcinogenic SF3b Mutations on Pre-MRNA Recognition Via All-Atom Simulations. *Biomolecules* **2019**, *9* (10). https://doi.org/10.3390/biom9100633.

(6) Wang, Q.; Hobbs, K.; Lynn, B.; Rymond, B. C. The Clf1p Splicing Factor Promotes Spliceosome Assembly through N-Terminal Tetratricopeptide Repeat Contacts. *J. Biol. Chem.* **2003**, *278* (10), 7875–7883. https://doi.org/10.1074/jbc.M210839200.

(7) Cretu, C.; Agrawal, A. A.; Cook, A.; Will, C. L.; Fekkes, P.; Smith, P. G.; Lührmann, R.; Larsen, N.; Buonamici, S.; Pena, V. Structural Basis of Splicing Modulation by Antitumor Macrolide Compounds. *Mol. Cell* **2018**, *70* (2), 265-273.e8. https://doi.org/10.1016/j.molcel.2018.03.011.

(8) Lange, O. F.; Grubmüller, H. Generalized Correlation for Biomolecular Dynamics. *Proteins Struct. Funct. Genet.* **2006**, *62* (4), 1053–1061. https://doi.org/10.1002/prot.20784.

(9) Galej, W. P.; Wilkinson, M. E.; Fica, S. M.; Oubridge, C.; Newman, A. J.; Nagai, K. Cryo-EM Structure of the Spliceosome Immediately after Branching. *Nature* **2016**, *537* (7619), 197–201. https://doi.org/10.1038/nature19316.

(10) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*. 1993, pp 779–815. https://doi.org/10.1006/jmbi.1993.1626.