Supplementary Information

Rewards of divergence in sequences, 3-D structures and dynamics of yeast and human spliceosome SF3b complexes

Arangasamy Yazhini, Sankaran Sandhya and Narayanaswamy Srinivasan*

Author Affiliations

Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Karnataka, 560012, India.

*Correspondence: Narayanaswamy Srinivasan, Molecular Biophysics Unit, Indian Institute of Science, Bangalore – 560012, Karnataka, India. ns@iisc.ac.in

Supplementary Text

S1. Distribution of SF3b6 across eukaryotes

Since the presence of SF3b6 distinguishes yeast and human SF3b complexes, we probed whether the absence of SF3b6 is specific to yeast. To recognize homologues in eukaryotes, SF3b proteins of these two organisms were queried against non-redundant RefSeq and Uniprot databases using BLASTP and HHblits algorithms, respectively (Camacho et al., 2009; Remmert et al., 2012). Reliable homologues were parsed using E-value (0.0001) and query coverage (70%) thresholds for both searches and 90% probability criterion for HHblits search, followed by careful manual evaluations. Search results show that SF3b6 homologues are present in 1308 species, which corresponds to 18 kingdoms, indicating that the protein is distributed across different lineages of eukaryotes (Table S4A). However, SF3b6 homologue could not be identified in 834 species that have homologues for at least one of the other SF3b proteins.

Although it is an arduous task to identify the absence of a protein in these species *in silico*, we have applied multiple criteria as described below to recognize species that potentially lack SF3b6. First, we have considered only 806 species of the fully sequenced genome for this study. Second, we carried out TBLASTN search against a recent version of the non-redundant nucleotide database. Here, we used a representative set of SF3b6 queries, obtained through clustering of identified SF3b6 homologues at 40% sequence identity (28 proteins) (Camacho et al., 2009). Based on the query coverage of 60% and E-value threshold of 10^{-12} criteria, we obtained a list of species in which homologues of SF3b6 could be recognized. These species names were then discarded from the consideration of the list of species that lack SF3b6. Finally, based on the premise that species with SF3b1 protein clasping SF3b6 binding domain would likely possess SF3b6, we ignored species having SF3b1 with SF3b6 binding domain from the list.

Through these filters, we have identified 215 species that potentially lack SF3b6 (Fig. S3A and Table S4B). Association of these species with the NCBI taxonomy tree of the entire species set covered in our dataset indicates that SF3b6 is universally absent in all members of Saccharomyces genus considered in this study. In other genera of Saccharomycetales order namely Candida, Eremothecium, Lachancea and Zygosaccharomyces, most members lack SF3b6 (Fig. S3A). In addition, among 17 Trypanosoma species covered in the analysis, SF3b6 is absent specifically in human infecting parasites *T. cruzi* and *T. brucei*. The absence of SF3b6 is also observed in other lineages of eukaryotes, including metazoans indicating that the loss of SF3b6 is common across eukaryotes. From this result, we infer that the role of SF3b6 in the SF3b complex is non-essential for 215 eukaryotic species. Hence, if one were to study the 3-D structure and function of SF3b complexes from these species, cryo-EM structures of yeast complex are suitable.

S2. Sequence conservation analysis of SF3b6 binding site in the SF3b1

SF3b1 is a binding partner for SF3b6 and hence the loss of SF3b6 potentially influences the evolution of SF3b1 in species having yeast-like SF3b complex. To study this aspect, we performed sequence conservation analysis for SF3b6 binding site in the SF3b1. As described in Supplementary Text S1, homologues of SF3b1 were collected from diverse eukaryotes. They were further clustered at 60% sequence identity using CD-HIT (Li and Godzik, 2006). This was done to minimize the dominance of densely populated close homologues on the sequence conservation profile. The final representatives of reasonably diverged homologues corresponding to 88 SF3b1 sequences were subjected to multiple sequence alignment using MAFFT algorithm (Katoh and Standley, 2013). Since SF3b1 homologues have conserved HEAT repeats, we used the L-INS-i option in the MAFFT algorithm, which is an iterative refinement protocol to generate a more accurate alignment for proteins having at least one region that can be aligned. For the interface conservation, alignment positions were selected using the human sequence as a reference. Figure S3B shows the sequence alignment of SF3b6 interface region in SF3b1 homologues from diverse eukaryotic lineages and the consensus sequence of 88 representative sequences obtained from cluster sets at 60% sequence identity. SF3b6 binding site is located in the N-terminal extension of the SF3b1. From the alignment, it is discernible that although SF3b6 interface region is conserved in several eukaryotic lineages, species lacking SF3b6 homologue such as yeast acquired extensive divergence (species highlighted in grey background in Fig. S3B). Especially, the N-terminal extension harbours deletion in these species suggesting that SF3b1 homologues have evolved substantially to compensate the loss of SF3b6 in the complex.

S3. Amino acid propensities at the interfaces of yeast and human SF3b complexes

The interfaces within a SF3b complex show species-specific interaction patterns in yeast and humans and are associated with sequence variations (Section 3.7 of Results and Discussion). This result prompted us to examine residue preferences at the interface between these two organisms. For this, we computed amino acid propensities for 12 interfaces formed within the human SF3b and 11 interfaces formed within the yeast homologue using the following formula,

$$P^{i}_{int} = N^{i}_{int} / N^{i}_{tot}$$

where P^{i}_{int} is the propensity of residue *i* at the interface region, N^{i}_{int} is the count of residue *i* at the interface region, and N^{i}_{tot} is the count of residue *i* present in the entire complex. N^{i}_{int} and N^{i}_{tot} are normalized by the total number of residues present at the interfaces and in the entire complex, respectively. The result shows that aspartate, glycine, isoleucine have no notable differences in their propensities for occurring in the interfaces between yeast and human SF3b complexes (Fig. S12). However, other residue types show altered interface propensities between these two organisms. We

find that cysteine, methionine, phenylalanine, threonine, tryptophan have higher interface propensities in humans, whereas proline and valine have higher propensities in yeast. This result indicates that residue preference at the interface regions differs between yeast and human SF3b complexes. We expect that such differences could contribute to species-specific interaction patterns at the interfaces, as observed in this study, and potentially modulate the interaction strength of inter-protein associations within the SF3b complex.

S4. Comparison of interface interaction energies between yeast and human SF3b complexes

To quantify the effect of altered residue propensities and interaction patterns on the interaction strength of protein partners, we compared the interface interaction energies between yeast and human SF3b complexes. We used B^{act} spliceosome assembly state structures (Protein Data Bank or PDB codes: 5GM6 for yeast and 5Z58 for humans) for this analysis. The structures were energy minimized and interaction energies of 12 interfaces in the human SF3b and 11 interfaces in the yeast homologue were calculated using FoldX suite (Schymkowitz et al., 2005). We find that the interaction energies of individual interface regions in human SF3b differ from equivalent interfaces in yeast SF3b by ~2kcal/mol to as high as 29kcal/mol (Table S6). For instance, the interaction energy of SF3b1 in complex with SF3b2 is -41.2kcal/mol, while the equivalent interface in the yeast homologue shows an interaction energy of -70.5kcal/mol. These energy differences are largely contributed by sidechain hydrogen bonds, van der Waals interactions and hydrophobic effect. Moreover, the interaction energy of SF3b3-SF3b1 is 3.4kcal/mol lower than the interaction between their homologues Rse1-Hsh155, indicating higher interaction strength in humans than yeast. The enhanced interaction strength in the human complex could result from additional interactions observed at the SF3b3-SF3b1 interface (Fig. 7C). Likewise, interaction energies of SF3b3-SF3b5 and Rse1-Ysf3 differ by 2.6kcal/mol that shows variations in local structures of interacting partners at the interface (Fig. 7D). Together, the interaction energy comparison reveals that the interaction strength between protein partners within the SF3b complexes varies substantially between yeast and humans.

S5. A survey of interacting protein partners for yeast and human SF3b proteins

Since, inter-protein interface regions which are otherwise known to be conserved show significant differences in the SF3b complexes, we reckon that the nature of interacting partners may have exerted an influence on the evolution of SF3b proteins in yeast and humans. To probe this, we surveyed interacting protein partners of SF3b proteins in the STRING database (Szklarczyk et al., 2019). We applied the following filters to retrieve reliable list of interacting proteins, i) the association should have an experimental evidence and ii) the combined score for a given association is above 900 *i.e.,* at

least 90% confidence. Based on these criteria, we collected a list of protein partners that interacts with any one of the SF3b proteins from yeast or humans and details are given in the Table S7. We find the number of interacting protein partners for yeast and human SF3b proteins vary remarkably. For instance, SF3b1 has about 235 versatile protein partners while its yeast homologue has 71 protein partners (Table S7A). Likewise, SF3b3 may interact with 30 additional protein partners compared to the yeast homologue Rse1. When we compare the nature of proteins and homologous relationships using BLASTp algorithm (Camacho et al., 2009), we observe that a considerable number of interacting partners are not common between yeast and humans (Table S7B). This observation suggests that yeast and human SF3b proteins have a distinct set of interacting proteins and may have been subjected to different evolutionary constraints.

Supplementary figures and legends

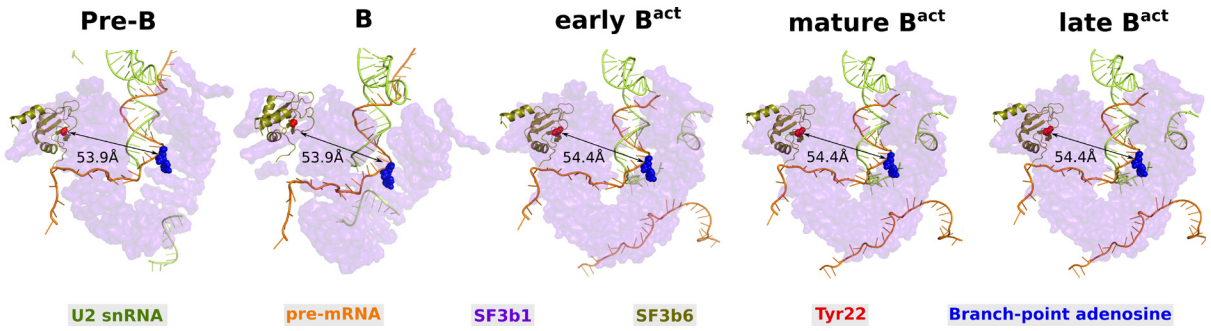


Fig. S1. Spatial distance between Tyr22 of SF3b6 and branch-point adenosine of pre-mRNA. Shown is the cartoon representations of SF3b1-SF3b6-U2 snRNA-(pre-) mRNA complex structures observed in Pre-B, B, early B^{act}, mature B^{act} and late B^{act} assemblies of the human spliceosome that are available as PDB entries 6AH0, 6AHD, 5Z58, 5Z56 and 5Z57, respectively. Molecular representation and color codes are as follow: U2 snRNA (ribbon, green); pre-mRNA (ribbon, orange); SF3b1 (surface, purple); SF3b6 (ribbon, olive); Tyr22 (sphere, red); branch-point adenosine (sphere, blue). The arrow shows the distance between the Cα atom of Tyr22 in SF3b6 and the phosphate atom of branch-point adenosine in pre-mRNA.

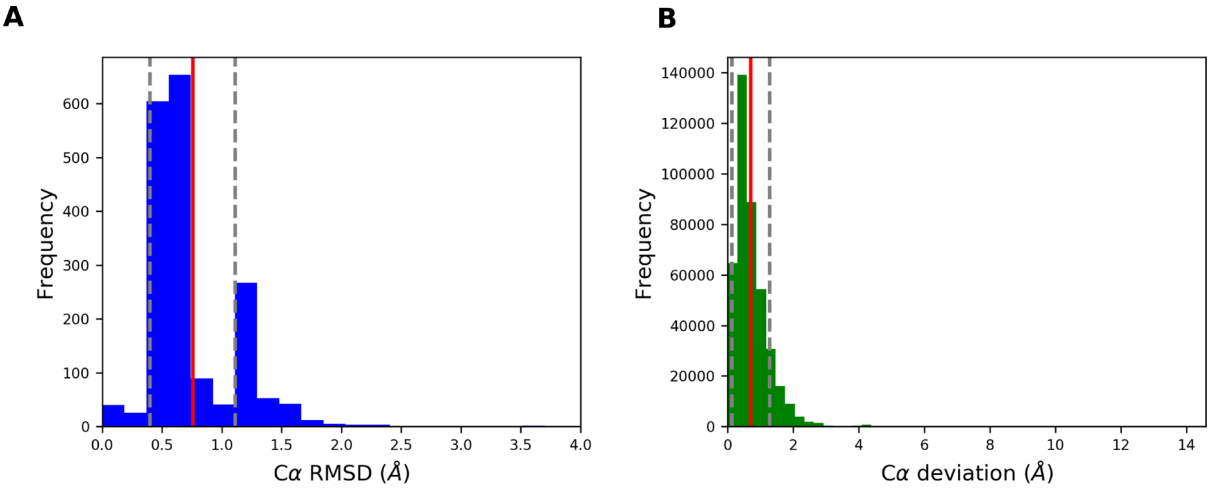


Fig. S2. Deviation in structures of the same proteins determined from different cryo-EM experiments. Shown are A) frequency distribution of Cα RMSD and B) frequency distribution of Cα distance between identical residue positions in different structures of the same protein. Mean and standard deviation values are indicated in red and grey lines, respectively.

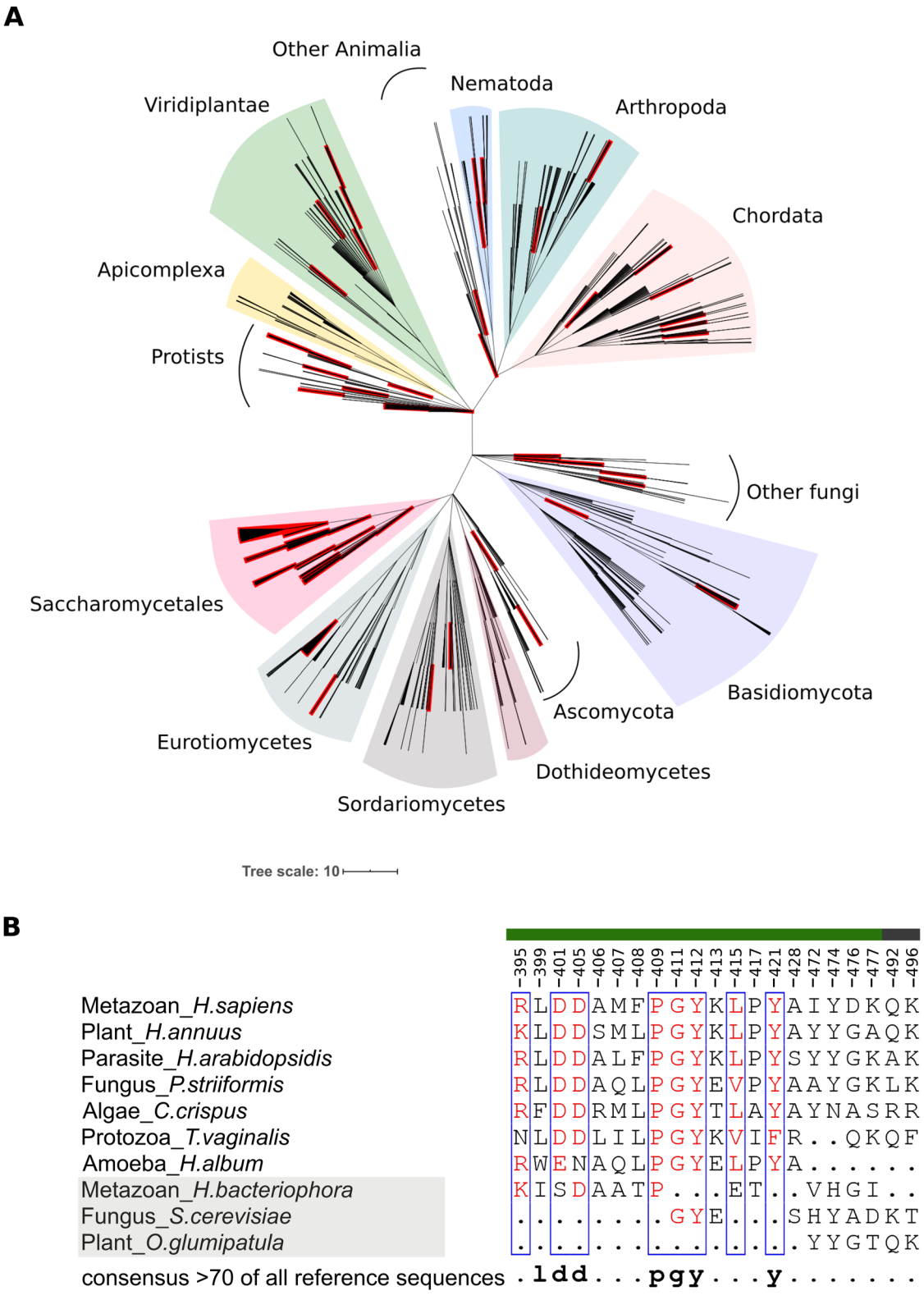


Fig. S3. Distribution of SF3b6 in eukaryotes and the binding site conservation. A) Shown is the NCBI taxonomy tree for 2142 species in which homologue of any one of the SF3b proteins could be identified. Red color on the node branch highlights species that lack SF3b6. Separate taxonomy clades are labelled by their names and highlighted by different background colors. The tree representation was generated using iTOL tool (Letunic and Bork, 2019). B) Sequence alignment of SF3b6 interface in

SF3b1. Shown are selected SF3b1 homologues from diverse eukaryotic lineages, including species in which SF3b6 is absent (highlighted in grey background). The alignment image was generated using ESpript (Robert and Gouet, 2014). The sequence label represents the species name (in italics) along with the name of its associated eukaryotic lineage. At the bottom, a consensus sequence from the alignment of 88 SF3b1 homologues is shown. The single letter residue codes indicate that the residue is physico-chemically conserved in 70% of the homologues. Green and grey bars at the top indicate the regions of N-terminal extension and HEAT repeats of SF3b1, respectively.

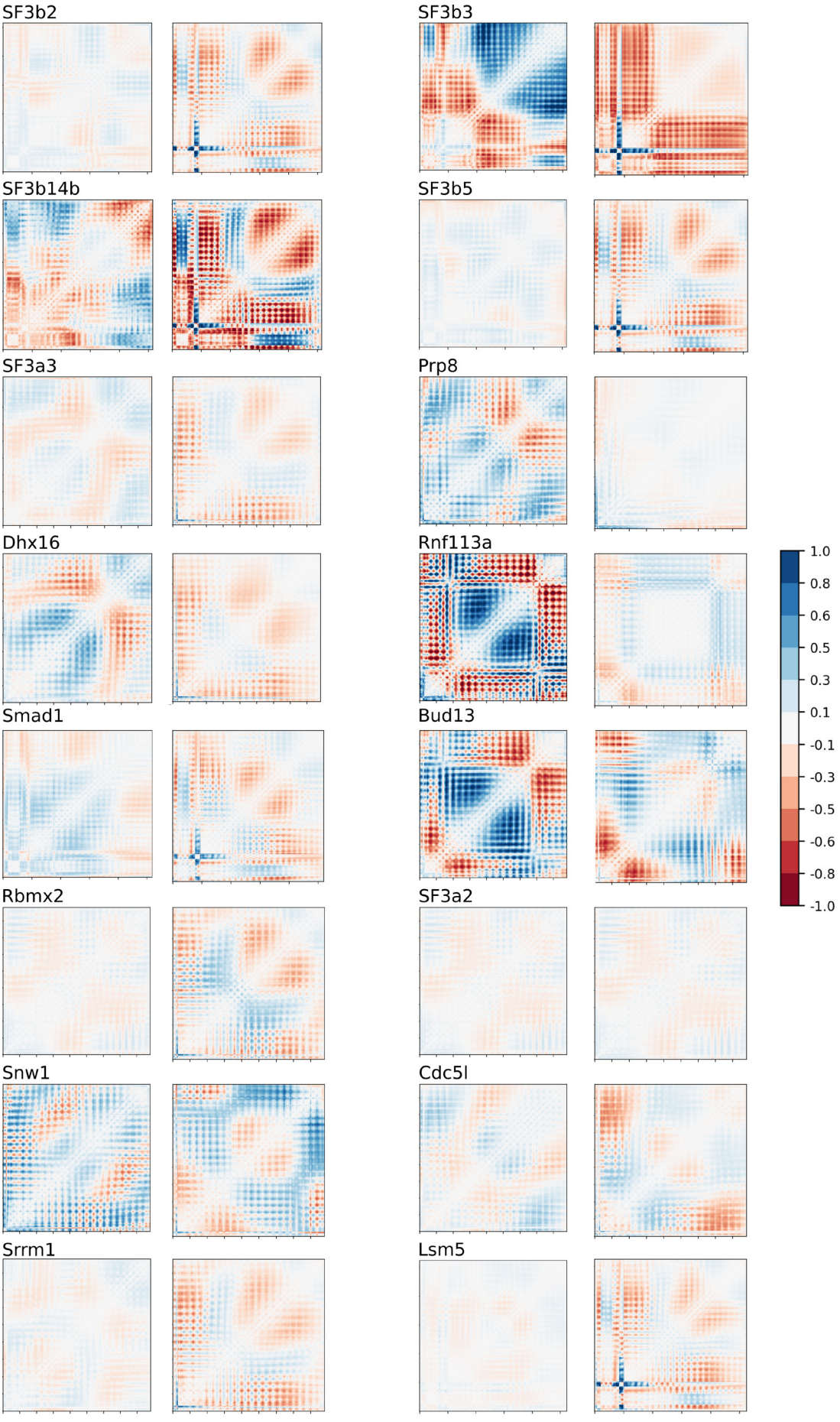

SF3b2
SF3b3
SF3b14b
SF3b5
SF3a3
Prp8
Dhx16
Rnf113a
Smad1
Bud13
Rbmx2
SF3a2
Snw1
Cdc5l
Srrm1
Lsm5
1.0
0.8
0.6
0.5
0.3
0.1
-0.1
-0.3
-0.5
-0.6
-0.8
-1.0

Fig. S4. Effect of SF3b6 on the intrinsic dynamics of SF3b1 in the presence of other spliceosome proteins. Paired heatmaps show difference in the cross-correlation matrices of SF3b1 in different contexts. In the left panel, the difference was calculated between a binary complex in which SF3b6 is bound (SF3b1-SF3b6) and a ternary complex (SF3b1-SF3b6-SF3b1 interacting protein). In the right panel, the difference was calculated between a binary complex in which SF3b1 interacting protein is bound (SF3b1-SF3b1 interacting protein) and a ternary complex (SF3b1-SF3b6-SF3b1 interacting protein). Denser color (correlation value > ±0.5) in the heatmap indicates that the presence of SF3b6 or SF3b1 interacting protein strengthens (positive value) or weakens (negative value) correlated motions of residues in the SF3b1. In total, 16 spliceosome proteins were analyzed (Table S5) and the result of each protein is labelled by their name.

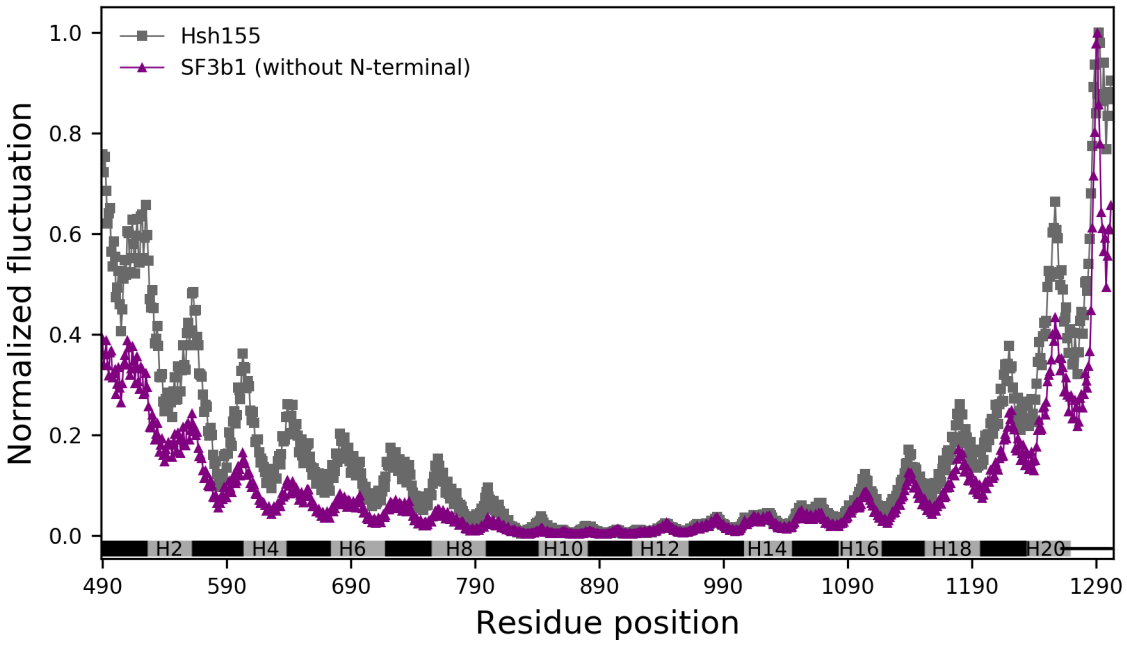


Fig. S5. Comparison of residue fluctuations between yeast Hsh155 (grey) and human SF3b1 (without N-terminal, magenta). Mean square residue fluctuations are normalized by the maximum residue fluctuation experienced in the protein anisotropic network model. HEAT repeat regions are labelled by numerical numbers.

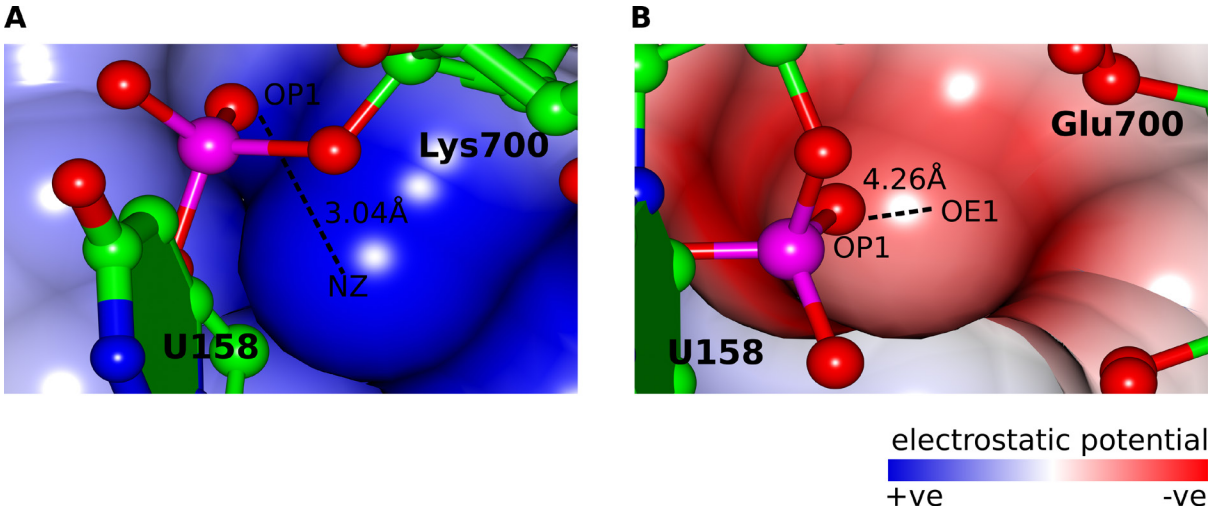


Fig. S6. Effect of Lys700Glu mutation on the pre-mRNA interaction of SF3b1. Shown in A) is the electrostatic potential surface (McNicholas et al., 2011) of SF3b1 protein with a focus on the Lys700 (PDB code: 5Z58). The inter-atomic distance of 3.04Å indicates that the phosphate ion of uracil base 158 in the pre-mRNA has a non-bonded interaction with the positively charged NZ atom of Lys700 in the SF3b1. B) Electrostatic potential surface of Glu700 variant of SF3b1 generated using *in-silico* mutagenesis in Chimera (Pettersen et al., 2004) and side-chain positions were optimized using SCWRL 4.0 algorithm (Krivov et al., 2009). Negatively charged oxygens in the carboxyl group of Glu700 create an electrostatically repulsive environment for the uracil base 158 of pre-mRNA.

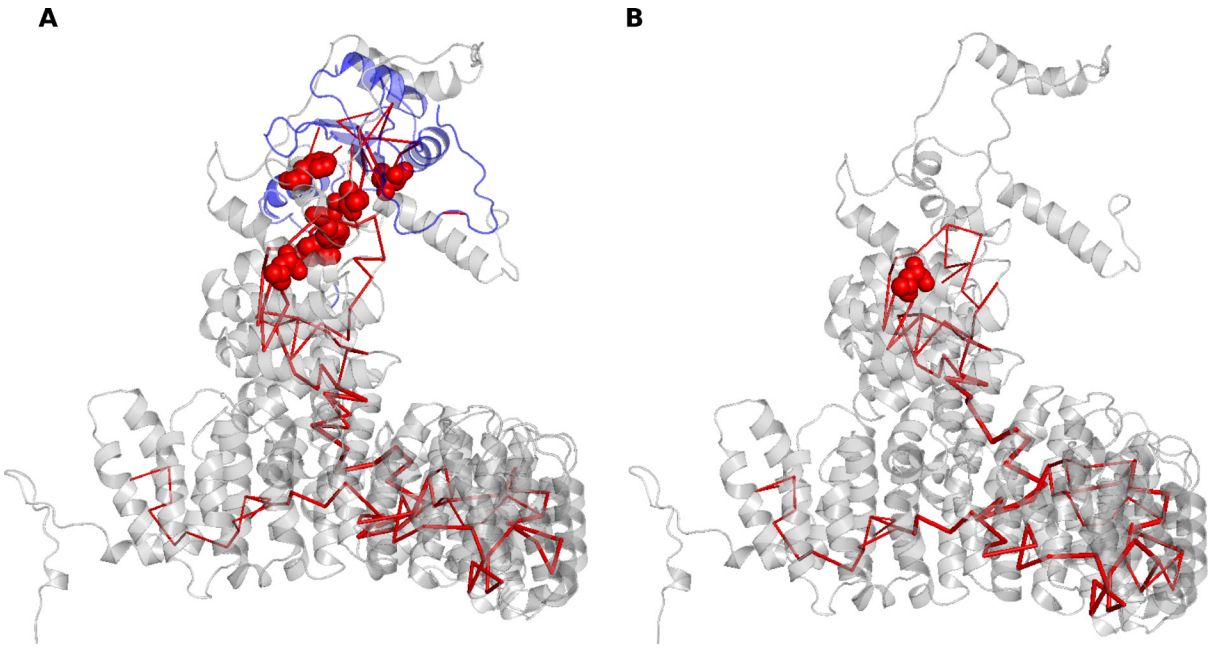


Fig. S7. Participation of SF3b6 interacting residues in the long-range residue-residue communication within SF3b1. A) A cartoon representation of SF3b1 (grey) bound to SF3b6 (blue)

along with red links showing the metapath that mediates long-range communication between two termini of HEAT repeats. Sphere representation highlights residues that interact with SF3b6 (Asn396, Phe408, Ile474, Tyr474 and Pro537) and pre-mRNA (Leu500). B) A cartoon representation of SF3b1 (grey) along with red links showing the metapath when SF3b1 is considered in isolation. Pre-mRNA binding residue (Leu500, shown as a sphere) is involved in the metapath, while SF3b6 interacting residues are not.

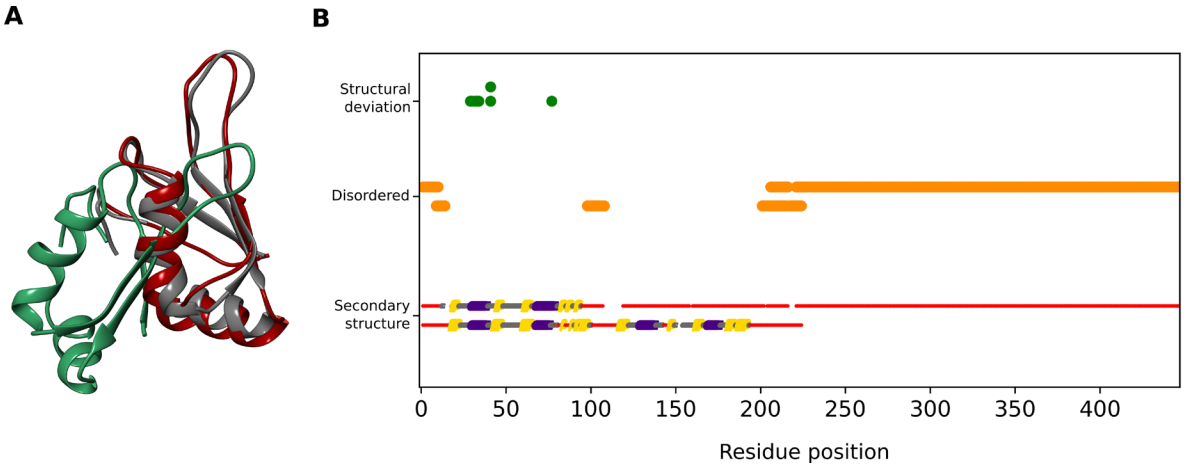
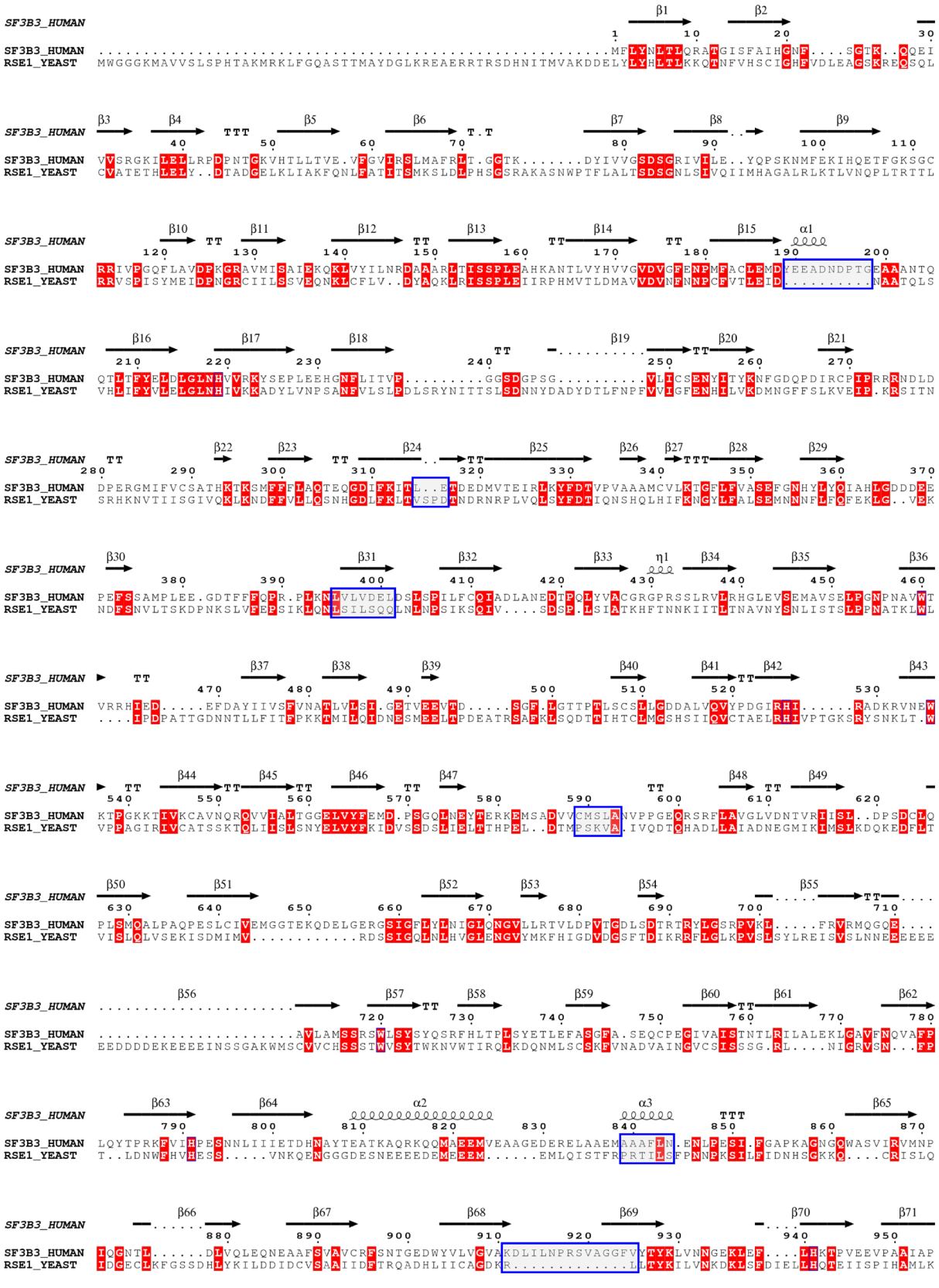


Fig. S8. Comparison of sequence and structural features between SF3b4 and Hsh49. A) A cartoon representation of superposed structures of SF3b4 (maroon) and Hsh49 (grey-superposed in isolation; green-superposed with the entire complex). B) The axis labelled as secondary structure shows the secondary structure conformation adopted by SF3b4 (top)/Hsh49 (bottom) in B^{act} spliceosome assembly. Missing regions are highlighted in red and gaps in the line plot indicate sequence insertions/deletions. In the axis with the label 'disordered', regions predicted to be disordered are highlighted (orange). In the axis labelled as structural deviation, green markers indicate structural variability observed between SF3b4 and Hsh49 upon superposition in isolation (top) and superposition along with the entire SF3b complex (bottom).


SF3B3_HUMAN
RSE1_YEAST

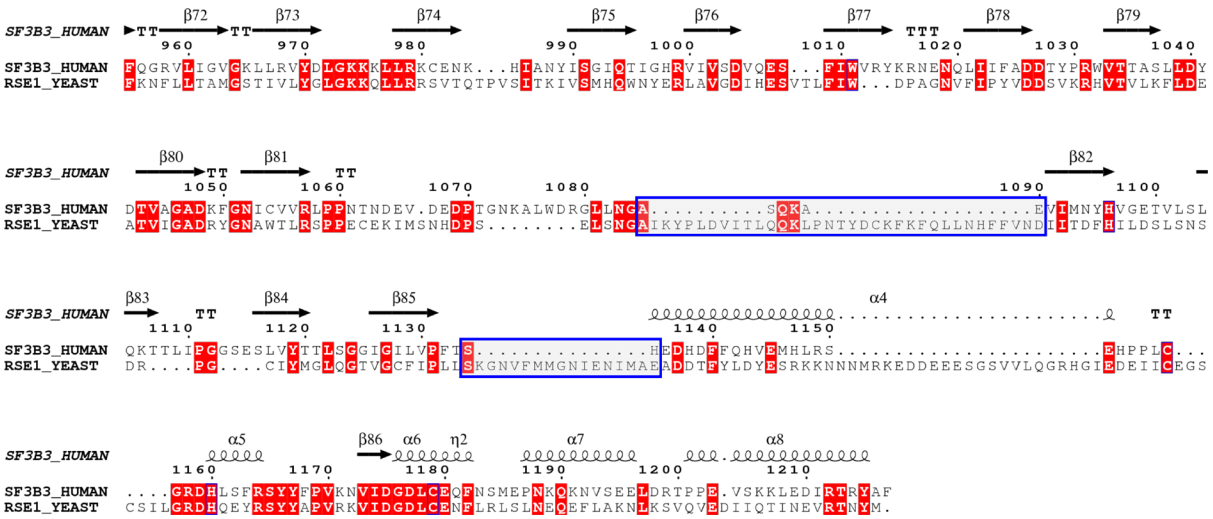


Fig. S9. Pair-wise sequence alignment between yeast Rse1 and human SF3b3. Regions with variability in the secondary structure conformation are highlighted (in blue box) and annotation on secondary structural information was obtained from the B^{act} structure of human SF3b3 (PDB code: 5Z58) (Robert and Gouet, 2014).

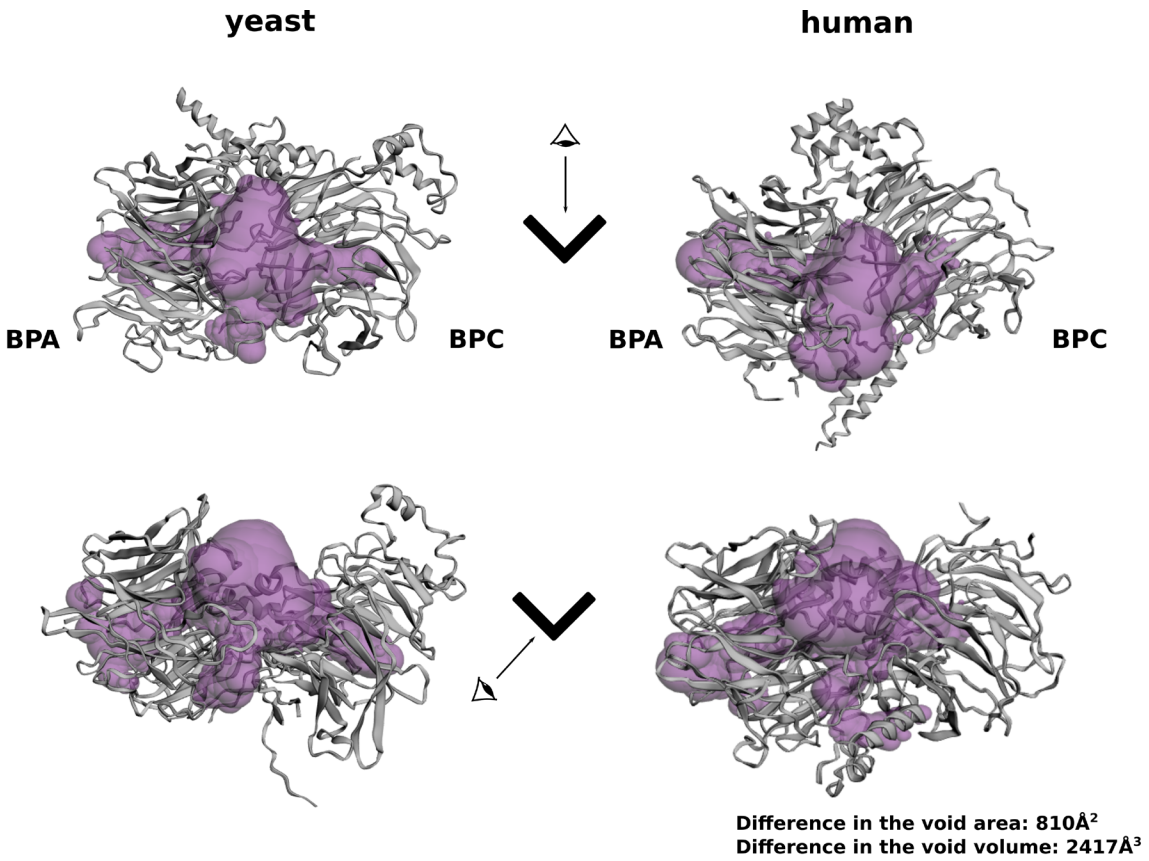


Fig. S10. The difference in the size of V-cleft cavity present at the junction of BPA and BPC domains of SF3b3/Rse1. The cavity is highlighted by surface representation in purple color.

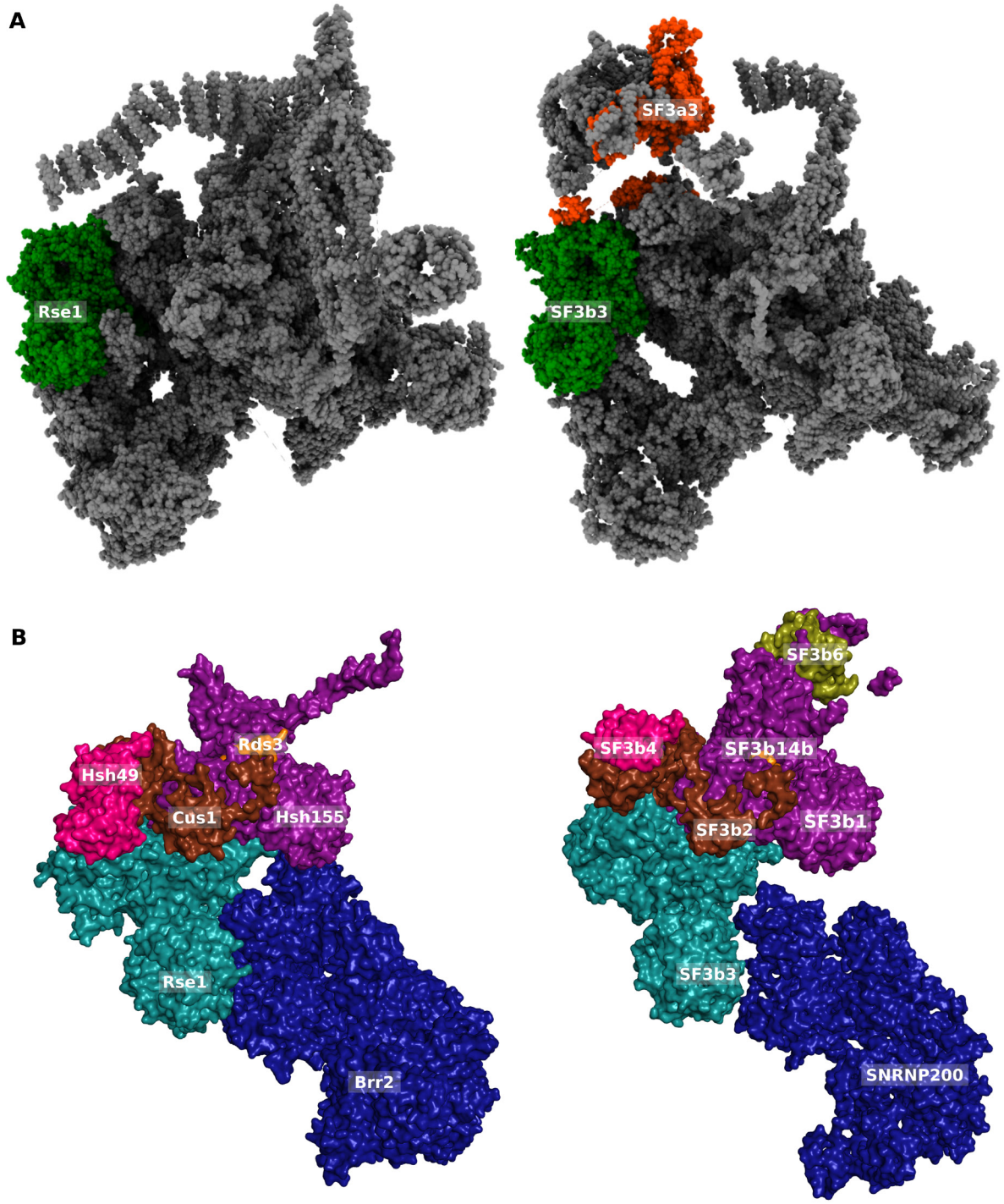


Fig. S11. Protein association around SF3b3/Rse1 in the B^{act} spliceosome assembly. A) Shown are the space filled representation of yeast and human B^{act} assembly structures with highlights on SF3b3/Rse1 (green) and SF3a3 (orange). B) Association of Brr2/SNRNP200 with the SF3b complex in yeast (left panel) and humans (right panel).

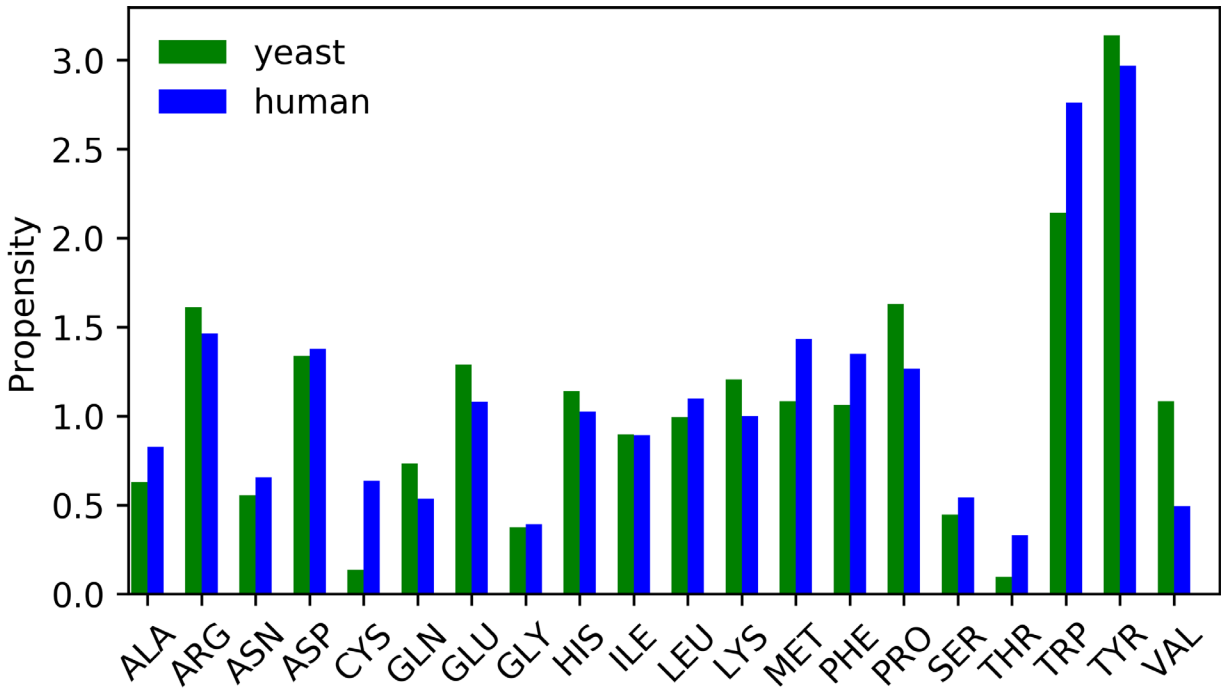


Fig. S12. Amino acid propensity at the interface regions of yeast and human SF3b complexes. Shown is the bar plot indicating propensity values of each residue type at the interface regions calculated for yeast (green) and human (blue) SF3b complexes.

References

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics *10*, 421.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Krivov, G.G., Shapovalov, M. V., and Dunbrack, R.L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins Struct. Funct. Bioinforma. *77*, 778–795.

Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res. *47*, W256–W259.

Li, W., and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

McNicholas, S., Potterton, E., Wilson, K.S., and Noble, M.E.M. (2011). Presenting your structures: The CCP4mg molecular-graphics software. Acta Crystallogr. Sect. D Biol. Crystallogr. *67*, 386–394.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. J.

Comput. Chem. *25*, 1605–1612.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175.

Robert, X., and Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res. *42*, W320-4.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: An online force field. Nucleic Acids Res. *1*, W382-388.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.