Spliceosomal proteins encoded by fungal genomes

Sandeep J. Sarde[1], Frank Kempken[1] and Abhishek Kumar[1,2,*]

[1]Abteilung Botanische Genetik und Molekularbiologie, Botanisches Institut und Botanischer Garten, Christian-Albrechts-Universität zu Kiel, Olshausenstr. 40 24098 Kiel, Germany
[2]Present address: Molecular Genetic Epidemiology (C050), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580 69120 Heidelberg, Germany

A large number of spliceosomal proteins are required for proper RNA splicing. While spliceosomal proteins from several model organisms have been analysed, only limited studies are available for fungal species. Hence, we have performed a comparative genomic analysis using eight fungal species belonging to three taxa (Ascomycetes, Basidiomycetes and Glomeromycota). We identified variable number of spliceosomal proteins in fungal species. From the small nuclear ribonucleoproteins (snRNPs), all the snRNPs were identified. In non-snRNPs, only some sub-groups were found extensively conserved in all fungal species, including PRP19 complex proteins, catalytic step II and late-acting proteins. In heterogeneous nuclear ribonucleoproteins (hnRNPs), variable number of proteins was identified. The number of spliceosomal proteins identified in filamentous fungi was higher than that in yeast. The collection of these spliceosomal proteins provides further insight into pre-mRNA splicing in fungi.

Keywords: Fungal genomes, pre-mRNA, snRNPs, spliceosomal proteins.

IN eukaryotes, genes are interrupted with non-coding sequences (introns), which are transcribed into pre-mRNA in the nucleus. The pre-mRNA is then processed and this results in the splicing out of introns to give yield a mature mRNA. This process, called splicing[1], serves as one of the hallmarks of eukaryotic genetics and is a crucial mechanism for eukaryotic messenger RNAs before they get translated into functional proteins[2]. This process is catalysed by the spliceosome, a multi-component macromolecular machine[3–5]. The spliceosome is a multi-megadalton ribonucleoprotein (RNP) complex comprising of pre-mRNA template, small nuclear ribonucleoproteins (snRNPs) and different non-snRNPs[1,6]. Further background of different groups of spliceosomal proteins is provided below and in Supplementary Section 1.

The snRNPs are core constituents of the spliceosomal complex which regulate pre-mRNA splicing and are comprised of a unique small nuclear ribonucleic acid (snRNA), a common set of Sm proteins (Sm-B⁄SmB', SmD1, SmD2, SmD3, SmE, SmF and SmG) and diverse number of snRNP specific proteins[7]. The U2 dependent spliceosome is assembled from the U1, U2, U5, U4/U6 snRNPs, and an abundant number of non-snRNP proteins. In contrast, the U12-dependent spliceosome is assembled from U11, U12, U5 and U4atac/U6atac snRNPs[6]. These different snRNPs are classified into Sm/LSm core proteins, U1, U2, U5, U4/U6 specific proteins and tri-snRNP specific proteins[8].

Sm/LSm proteins are present ubiquitously in eukaryotes. They interact with RNAs to make complexes, thus taking part in nearly every cellular process. Due to structural similarity with Sm proteins, they are called 'Like Sm' or 'LSm' proteins[9]. Sm proteins are a set of small polypeptides that play a critical role in gathering the U1, U2, U5 and U4/U6 snRNPs for pre-mRNA splicing[10]. Sm proteins are differentiated into seven sub-classes based on human Sm proteins, which are known as SmB, SmD1, SmD2, SmD3, SmE, SmF and SmG. There are a total of nine LSm proteins (LSm1 to LSm9) existing in *S. cerevisiae*, of which, LSm2 to LSm7 appear to be very similar to SmD1 to SmG respectively[11]. In contrast, LSm1 and LSm8 seem to be more similar to the SmB sub-family[11,12].

The U1 snRNP is a vital member of the spliceosomal snRNPs. Human U1 snRNP consists of several specific and unique snRNPs including 164-nucleotide U1 small nuclear RNA (U1SnRNA)[13]. In metazoans, splicing mechanism is initiated by U1snRNA (part of U1 snRNP), by recognizing 5′-splice-site (5′-ss) and forming the E-complex. Subsequently, this base pairing of U1RNA is stabilized by U1 snRNP specific proteins named U1-70K and U1-C[7].

U2 snRNP firmly associates with the branching site[14] after the U1 snRNP, forming pre-spliceosomal complex A or pre-spliceosome complex. U2 snRNP plays a main part in splicing after the dissociation of U1 and U4 snRNP from pre-mRNA. A wide ranging base pairing system and conformational changes are shaped between U6 and U2 snRNP, which juxtaposes the branch site (BS) and 5′-ss for the initial step of splicing[6].

Many eukaryotic genes are expressed as precursor mRNAs by RNA polymerase II, which are further

*For correspondence. (e-mail: abhishek.abhishekkumar@gmail.com)

processed into mature mRNA by splicing. These mRNA precursors primarily produced by RNA polymerase II are accompanied by proteins in large complexes[15], and collectively termed as hnRNPs[16].

The hnRNPs encompass a family of RNA-binding proteins, which is very complex and diverse. It is associated with several functions like processing heterogeneous nuclear RNAs (hnRNAs) into mature mRNAs or acting as trans-factors in regulating gene expression[15]. In the nucleus, they primarily participate in RNA splicing[17], and transcriptional regulation[18]. Moreover, some members of the hnRNP family are required for alternative splicing. For example, hnRNP A1 can regulate 5′-ss selection with the help of splicing factor SF2 (ref. 19).

Previous studies on spliceosomal proteins have focused solely on a limited number of model organisms like *Homo sapiens*, *Arabidopsis thaliana*, flies, yeast and *Dictyostelium discoideum*[8]. The available data on these model organisms have improved the understanding of splicing mechanism on a molecular level. However, very little is known about pre-mRNA splicing in fungi with the exception of baker's yeast (*Saccharomyces cerevisiae*). Herein, we present a comparative genomic analyses of fungal spliceosomal proteins using eight different species.

Material and methods

We scanned eight fungal genomes (Table 1) for their putative spliceosomal proteins using different homology detections (Supplementary Figure 1) as described in the Supplementary Section 2. Pfam domain and phylogenetic analyses were performed using HMMER 3.0 (ref. 20) under CLC bio genomics workbench 7.5 (www.clcbio.com) and MEGA6 tool[21] respectively. A comprehensive approach is provided in the Supplementary Section 2.

Results

Status of spliceosomal snRNP proteins

There are a total of 49 human snRNP proteins, which are categorized into different sub-groups like U1, U2, U5, U4/U6, U4/U6.U5 specific proteins and Sm/LSm proteins[8]. All sequences of these 49 human snRNP proteins were used to screen the proteomes of eight different fungal species.

Summary of Sm/LSm core proteins

The sub-group of Sm/LSm core proteins comprises 15 proteins, 7 in the LSm family and 8 in the Sm family[7]. We identified at least one homologue in all fungi analysed (Table 2). All the studied fungal species led to similar orthologs when searched with either the human SmB/B′ or SmN proteins (Table 2 and Figure 1 *a*).

The Sm/LSm proteins in the studied fungal species showed a one-to-one relationship with their human counterparts, except for some proteins in *N. crassa*, *R. irregularis* and *R. solani*. These proteins include SmD1, LSm2, LSm5 and LSm8, which were found to be duplicated when compared to their single human counterparts. Two homologues of human LSm7 and LSm6 proteins were also identified in *R. irregularis* and *R. solani* respectively. The two copies of LSm6 protein found in *R. solani* were 100% identical. The close one-to-one relationship of LSm3 and LSm4 proteins between human and fungal species from different phyla is illustrated in the phylogenetic tree (Figure 1 *b*).

Pfam domain analysis with these core proteins clearly depicts the predominant presence of a LSM (PF01423) domain. LSM (like Sm) have diverse functions, and are key regulators of RNA biogenesis and splicing[10].

Overview of U1, U2 and U5 snRNPs specific proteins

There are six, twelve and eight different proteins present in human U1, U2 and U5 snRNPs respectively. Searches employing human U1-snRNPs gave rise to a one-to-one relationship between human and different fungal species (Table 3) with the exception of the gene encoding CROP protein, which is duplicated in *R. irregularis*.

Pfam domain analysis with U1 snRNPs revealed that every protein from this group has a combination of known protein domains (Figure 2 *a*), which are crucial for splicing. SNRP70 protein has two domains: U1snRNP70 (Pfam domain ID. PF12220) and RNA recognition motif (RRM) (PF00076). Both these domains are often found in association. U1 snRNP70 wraps around the core domain of U1 snRNP and eventually interacts with U1-C, which is important for 5′-ss recognition[22]. SNRPC protein consists of zf-U1 domain (PF06220), a domain largely found in U1 which binds to the pre-mRNA 5′-ss at the initial stages of spliceo some assembly[23]. SNRPA protein consists of a single RRM domain (PF00076). This RRM domain is found in several RNA binding proteins, including protein components of snRNPs involved in regulating alternative splicing and hnRNP proteins[24].

In U2 snRNP, several proteins have a one-to-one relationship between human and the studied fungal species except for SF3b49 (Hsh49p), SF3a66 (Prp11p) and SF3b130 (Rse1p). These three proteins are duplicated in the *N. crassa* genome. Interestingly, there were also some proteins from this group not identified in some fungal species like SF3a60 (Prp9p), SF3b49 in *R. irregularis*, SF3b14b (Rds3p) in *R. solani* and SF3b10 (na), SF3b14a (na) in *S. cerevisiae* (Table 4).

Table 1. Summary of fungal genomes used in this study

Fungal strain	Abbreviation	Genome size (Mb)	No. of genes	Disease	Reference
Ascomycetes					
Aspergillus niger ATCC 1015 v4.0	Ani	34.85	11,910	*A. niger* causes black mould diseases on set of fruits and vegetables	http://jgi.doe.gov/fungi
Fusarium graminearum v1.0	Fgr	36.45	13,322	*F. graminearum* is plant pathogen and it causes Fusarium head blight to wheat and barley	www.broadinstitute.org/
Neurospora crassa OR74A v2.0	Ncr	41.04	10,785	–	www.broadinstitute.org/
Pyronema confluens CBS100304	Pco	50.03	13,367	–	Traeger *et al.*
Saccharomyces cerevisiae M3707	Sce	11.51	5,974	*S. cerevisiae* causes rarely uncommon human infection, known as as *S. cerevisiae fungemia*	Brown *et al.*
Basidiomycetes					
Rhizoctonia solani AGI-1B	Rso	47.66	15,157	*R. solani* is plant pathogen and it is causative agents of collar rot, root rot, damping off and wire stem	Wibberg *et al.*
Botryobasidium botryosum v1.0	Bbo	46.67	16,526	–	Riley *et al.*
Ustilagomaydis	Uma	19.68	6,522	*Ustilagomaydis* is also plant pathogen, which causes mut on maize and teosinte	www.broadinstitute.org/
Glomeromycota					
Rhizophagus irregularis DAOM 181602	Rir	91.08	30,282	–	Tisserant *et al.*

Table 2. Summary of Sm/LSm core proteins in selected fungi

Human proteins (Acc nos.)	Protein name[$]	HUGO[#]	Ascomycetes				Basidiomycetes			Glomeromycota	Yeast
			NCr	*Pco*	*Ani*	*Fgr*	*Bbo*	*Rso*	*Uma*	*Rir*	*Sce*
NP_003082	Sm B/B (Smb1p)	SNRPB	*673*	*7942*	*44190*	*1270*	*126879*	*75*	*3889*	*344619*	*37244*
NP_008869	Sm D1 (Smd1p)	SNRPD1	2231 2230	12496	1147070	444	48877	3931	3142	300907	33818
NP_808210	Sm D2 (Smd2p)	SNRPD2	7349	10659	1141771	8113	129639	1620	4781	333900	35689
NP_004166	Sm D3 (Smd3p)	SNRPD3	7192	1324	1126637	6917	25686	13256	–	51923	38624
NP_003086	Sm F (Smx3p)	SNRPF	3383	5532	1173444	1605	36592	1537	–	336789	32420
NP_003087	Sm G (Smx2p)	SNRPG	1678	1369	1186221	895	101193	1859	1525	341255	37714
NP_003088	Sm N (Smb1p)	SNRPN	*673*	*7942*	*44190*	*1270*	*126879*	*75*	*3889*	*344619*	*37244*
NP_067000	LSm2 (Lsm2p)	LSM2	1630 1629	12282	–	446	26309	–	3799	345622	33209
NP_055278	LSm3 (Lsm3p)	LSM3	3785	4045	1149216	3127	31768	337	1249	166862	35913
NP_036453	LSm4 (Lsm4p)	LSM4	5739	9185	1144560	6174	34888	3899	4446	176309	37110
NP_036454	LSm5 (Lsm5p)	LSM5	7333 7332	11001	1100652	7881	53821	921	190	340238	38391
NP_009011	LSm6 (Lsm6p)	LSM6	996	2733	1147756	555	115026	78794182	892	347097	37026
NP_057283	LSm7 (Lsm7p)	LSM7	516	6598	1142689	9257	106202	2761	2355	65468 339396	34661
NP_057284	LSm8 (Lsm8p)	NAA38	1118 1117	1219	1188730	564	125003	1484	–	339425	37527
NP_003085	SmE (Sme1p)	SNRPE	5898	4439	1143812	3140	169179	196	2679	79969	31311

Italic protein ids with underline mention to sequence being used two or more times. '–' just indicates that no hit is obtained/identified in the respective organism and is applicable for all the tables. [$]Conventional protein name with the yeast homolog is included in the parenthesis. [#]HUGO-gene symbol approved by Human Genome Organization Gene Nomenclature Committee. Underlined gene symbol is not approved.

Table 3. Comparative status of U1-snRNP specific proteins in selected fungi

Human proteins (Acc nos.)	Protein name[$]	HUGO[#]	Ascomycetes				Basidiomycetes			Glomeromycota	Yeast
			NCr	*Pco*	*Ani*	*Fgr*	*Bbo*	*Rso*	*Uma*	*Rir*	*Sce*
NP_003084	U1 C (Yhc1p)	SNRPC	7771	5649	35817	11732	31767	2258	5608	338665	35720
NP_004587	U1 A (na)	SNRPA	*6261*	*5949*	*1188131*	*3744*	*32694*	*959*	*807*	*334083*	*36013*
NP_003080	U1-70K (Snp1p)	SNRP70	4930	5119	1131062	7747	74947	2952	4775	76176	36126
NP_006098	CROP (Luc7p)	LUC7L3	8865	10167	1187737	2508	104804	14173	2646	67848 71989	30378
NP_060362	Fnbp3 (Prp40p)	PRPF40A	*1091*	*2489*	*1188615*	*8210*	*126520*	*3591*	*5364*	*230875*	*35136*
NP_001026868	HYPC (Prp40p)	PRPF40B	*1091*	*2489*	*1188615*	*8210*	*126520*	*3591*	*5364*	*230875*	*35136*

Italic protein ids with underline mention to sequence being used two or more times. '–' just indicates that no hit is obtained/identified in the respective organism and is applicable for all the tables. [$]Conventional protein name with the yeast homolog is included in the parenthesis. [#]HUGO-gene symbol approved by Human Genome Organization Gene Nomenclature Committee. Underlined gene symbol is not approved.

Table 4. Overview of U2-snRNP specific proteins in selected fungi

Human proteins (Acc nos.)	Protein name[$]	HUGO[#]	Ascomycetes				Basidiomycetes			Glomeromycota	Yeast
			NCr	*Pco*	*Ani*	*Fgr*	*Bbo*	*Rso*	*Uma*	*Rir*	*Sce*
NP_003083	U2 B (Msl1p)	SNRPB2	*6261*	*5949*	*1188131*	*3744*	*32694*	*959*	*807*	*334083*	*36013*
NP_005841	SF3b49	SF3B4	8179 8178	9007	1119561	8507	28940	12245	4679	–	31094
NP_005868	SF3a120 (Prp21p)	SF3A1	536	9773	1112452	9253	104317	2408	305	22885	36532
NP_006793	SF3a60 (Prp9p)	SF3A3	7654	7798	1215424	6797	165392	2543	6130	–	30459
NP_057131	SF3b14a (na)	SF3B14	2452	9176	1145176	643	69514	2548	2609	337327	–
NP_112577	SF3b10 (na)	SF3B5	439	3516	1147519	9397	159309	4686	748	262489	–
NP_116147	SF3b14b (Rds3p)	PHF5A	4664	3680	1159844	7799	33935	–	4493	340349	32540
NP_009096	SF3a66 (Prp11p)	SF3A2	6390 6389	7121	1182142	1387	582622	10275	5859	33865	30448
NP_003081	U2 A′ (Lea1p)	SNRPA1	5640	179	1099943	6156	26672	285	4654	171194	32963
NP_036565	SF3b155 (Hsh155p)	SF3B1	2257	4357	1121816	8153	192791	2765	5241	343790	31789
NP_036558	SF3b130 (Rse1p)	SF3B3	4716 4717	12472	1129671	8392 6102	26066	14327	737	341744 177723	32278
NP_006833	SF3b145 (Cus1p)	SF3B2	3017	11380	1161381	3707	164137	3004	5454	65618	31866

Italic protein ids with underline mention to sequence being used two or more times. '–' just indicates that no hit is obtained/identified in the respective organism and is applicable for all the tables. [$]Conventional protein name with the yeast homolog is included in the parenthesis. [#]HUGO-gene symbol approved by Human Genome Organization Gene Nomenclature Committee. Underlined gene symbol is not approved.

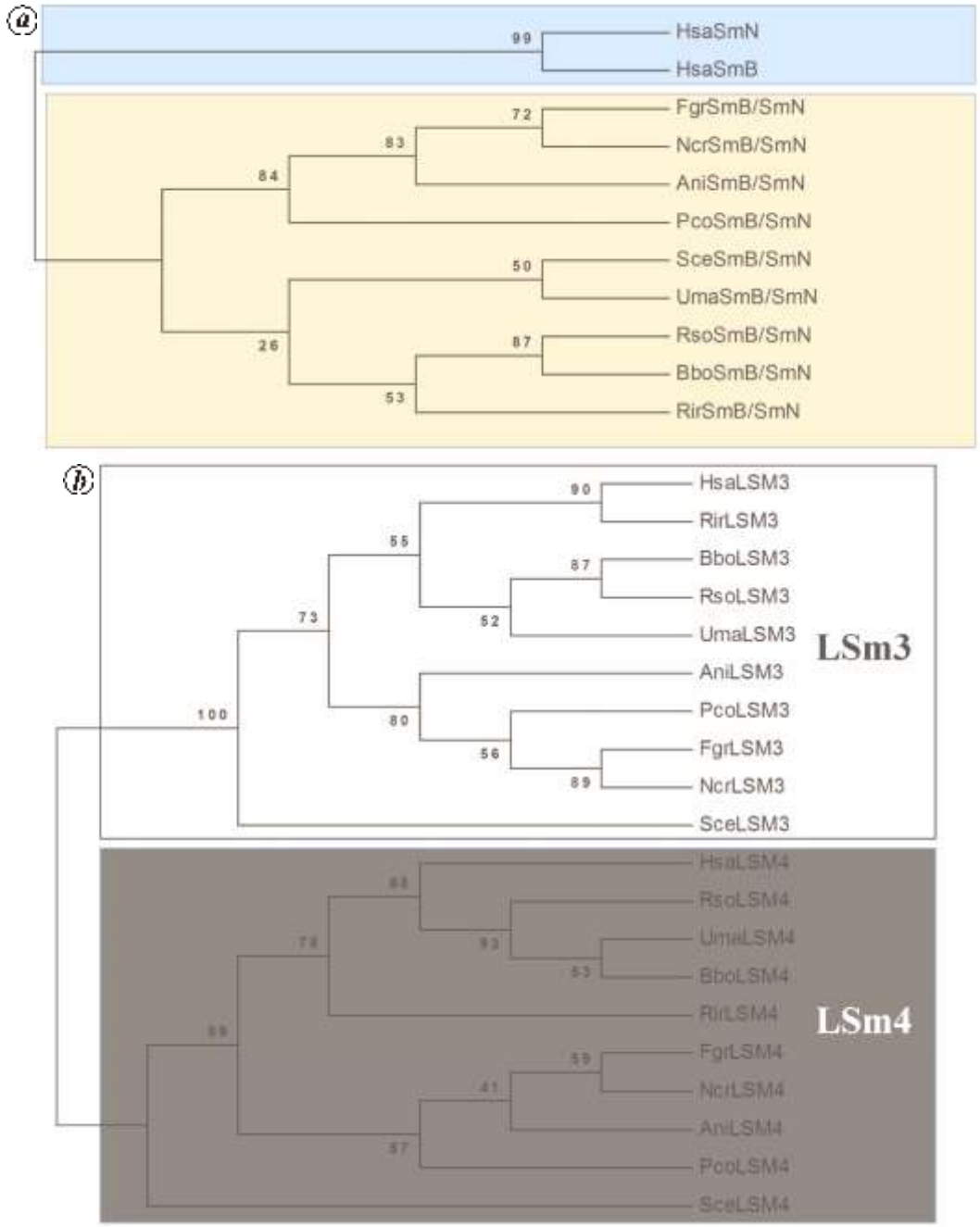

Figure 1. Phylogenetic analyses of Sm/LSm proteins of human and several fungal species of different phyla. ***a***, Evolutionary history of fungal SmB (SNRPB) and SmN (SNRNP). One-to-one orthologs of these two human genes SmB (SNRPB) and SmN (SNRNP) were detected in the studied fungal genomes. ***b***, Phylogenetic tree of LSm3–LSm4 proteins. One copy each of these two genes were detected in fungal genomes.

Pfam domain analyses of the U2 snRNP show that the RRM domain frequently occurs in proteins of different species. The RRM domain is found in several proteins of this group like SNRPB2, SF3B4 and SF3B14 (Figure 2 *b*). The SNRPA1 protein consists of a leucine rich repeat (LRR) domain (PF14580), which is often involved in the formation of protein–protein interactions[25]. The SF3A1 protein comprises two domains: PRP21 (PF12230) and Surp (PF01805), which are consistently seen in association. The PRP21 (PF12230) domain of SF3A1 protein plays a critical role in the assembly of the pre-spliceosome, whereas the Surp (PF01805) domain is found in regulatory proteins of splicing[26]. The SF3a3 protein comprises three domains, namely SF3a60 (PF12108), zf-C2H2 (PF12171) and DUF3449 (PF11931). The SF3a60 domain is essential for splicing and participates in forming the SF3a complex by providing a binding site for SF3a120 (ref. 27). Likewise, SF3B2 protein

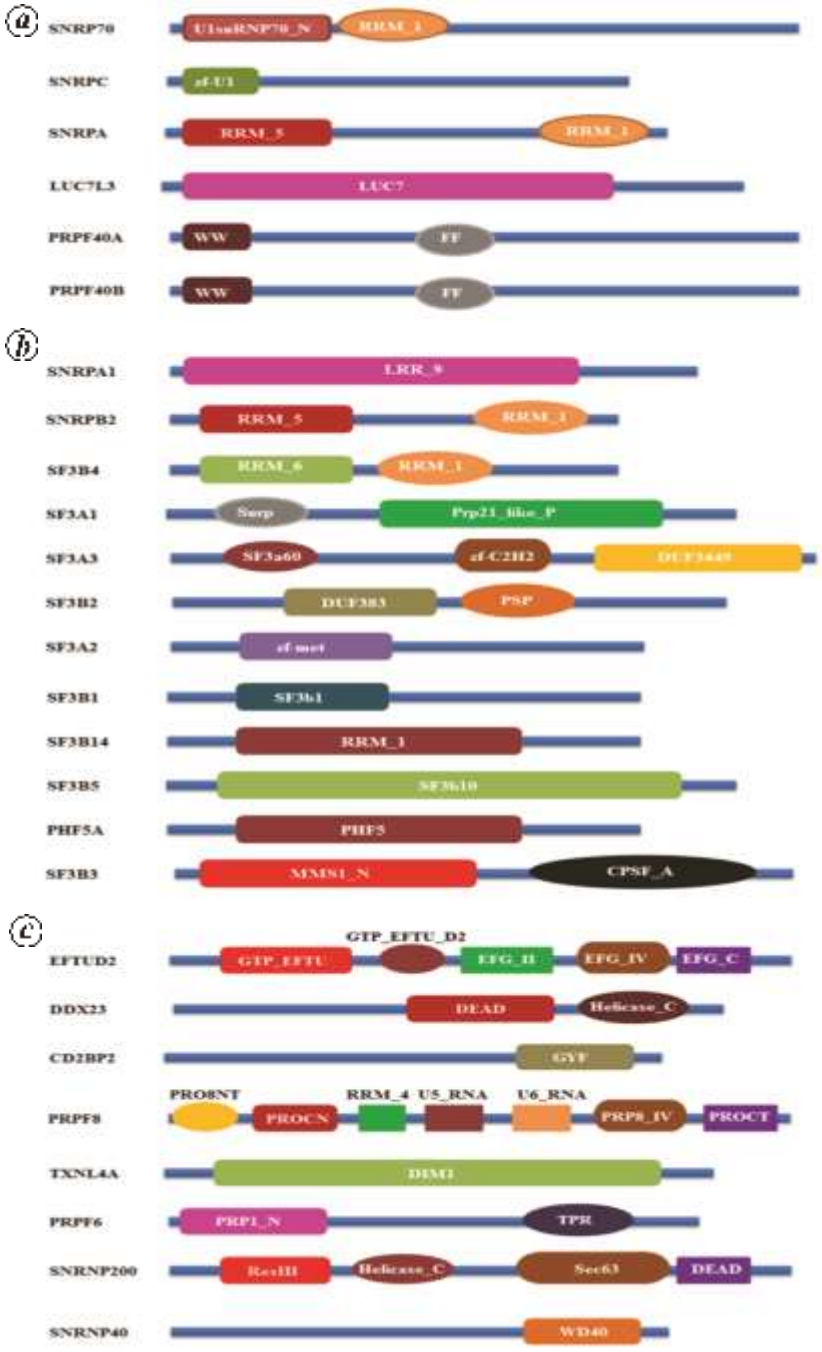

Figure 2. Overview of Pfam protein domains identified in different groups of snRNPs in human and different fungal genomes. ***a***, U2 snRNPs; ***b***, U5 snRNPs; ***c***, U1 snRNPs.

contains two domains, PSP (PF04046) and DUF382 (PF04037). PSP is a proline-rich domain of unknown function found in spliceosome-associated proteins, while DUF382 domain is specific to the human splicing factor 3b subunit 2 (ref. 28). SF3A2 protein consists of single zf-met (PF12874) domain. This zf-met domain is found in multiple copies in several proteins from plants to metazoans and considered to be an RNA-binding domain. Similarly, SF3B5 protein consists of a single SF3b10 (PF07189) domain, a subunit of splicing factor SF3b. SF3b associates with splicing factor SF3a and a 12S RNA unit to form the U2 snRNP complex essential for splicing[28]. The PHF5A protein contains a PHF5 (PF03660) domain. PHF5 belongs to the superfamily of PHD-finger proteins and is vital for splicing[29]. Moreover, SF3B3 protein comprises two domains named MMS1 (PF10433) and CPSF_A (PF03178). MMS1 is reported to protect against replication-dependent DNA damage in *S. cerevisiae*[30], whereas CPSF domain is required for splicing of single-intron pre-mRNAs[31]. A number of proteins of the U5

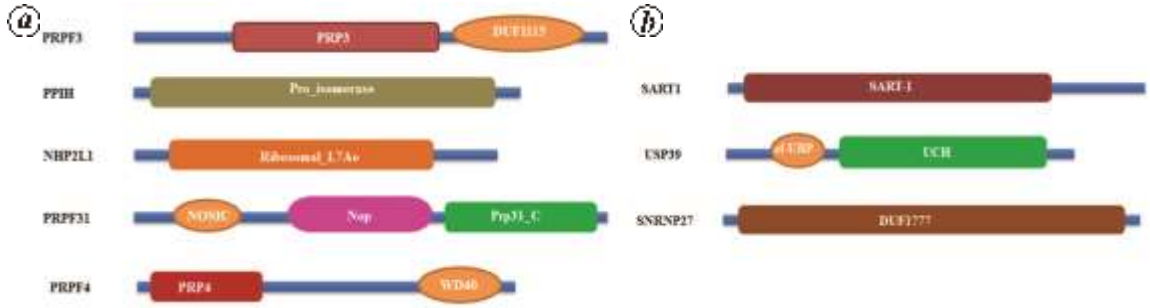

Figure 3. Summary of Pfam protein domains present in different groups of snRNP proteins in human and different fungal genomes. ***a***, U4/U6 snRNPs; ***b***, U4/U6.U5 tri-snRNP.

snRNP sub-group have shown similar one-to-one relationship like U1 and U2 snRNP proteins, except few proteins like U5–15 kDa or/and U5–200 kDa. These two proteins have paralogs in *N. crassa* and *R. irregularis*.

However, a few proteins were not identified in some of the fungal genomes like U5–52 kDa in *U. maydis* or and U5–40 kDa in *S. cerevisiae* (Supplementary Table 2). The fungal proteins of this group have similar domains as their human counterparts (Figure 2 *c*). The CD2BP2 (U5–52 kDa) protein has shown a single GYF domain (PF02213) identified in human and in the studied fungal species. The GYF domain is found in several eukaryotic proteins and it has been proposed that it may play a role in proline-rich sequence recognition[32,33]. The TXNL4A (U5–15 kDa) protein possesses a DIM1 (PF02966) domain, which has been identified in human and all studied fungal species. The human TXNL4A protein contains 37 extra residues that form putative binding sites for other spliceosomal factors[34]. Moreover, four domains were identified (Figure 2 *c*) in the SNRNP200 (U5–200 kDa) protein in humans and all the analysed fungal species, which include Sec63 (PF02889), DEAD (PF00270), ResIII (PF04851) and Helicase (PF00271). The Sec63 domain is essential for the assembly of functional endoplasmic reticulum translocons[35,36]; whereas in yeast this domain is found in pre-mRNA splicing proteins. The DEAD domain is involved in several aspects of RNA metabolism[37,38]. In the PRPF6 (U5–102 kDa) protein, two domains named PRP1 (PF06424) and TPR (PF13181) were identified. PRP1 is involved in mRNA splicing, RNA nuclear export and cell cycle progression[39], whereas the TPR (tetratricopeptide repeat) domain is a structural motif found in several proteins[40].

In the DDX23 (U5–100 kDa) protein, DEAD and Helicase domains were conserved in humans and included fungi (Figure 2 *c*). The PRPF8 protein consists of seven Pfam domains, which include PROCN (PF08083), PRP8 (PF12134), U6-snRNA (PF10596), PRO8NT (PF08082), U5 snRNA (PF10597), PROCT (PF08084) and RRM (PF10598) domains. PROCN is the vital domain in the PRO8 family and involved in pre-mRNA splicing[41]. PRP8 is a selenomethionine domain and assumed to be interacting with the spliceosomal core[42]. In the SNRNP40 (U5–40 kDa) protein, a WD40 domain was identified throughout all the analysed species. The EFTUD2 (U5–116 kDa) protein possesses five domains, which include GTP_EFTU (PF00009), EFG_IV (PF03764), EFG_C (PF00679), EFG_II (PF14492) and GTP_EFTU_D2 (PF03144) in the analysed species.

Summary of U4/U6 di-snRNP, U4/U6 and U5 tri-snRNP specific proteins

There are five and three specific proteins belonging to U4/U6 di-snRNP and U4/U6.U5 tri-snRNP sub-groups respectively[8]. Three among five U4/U6 di-snRNPs have counterparts in all the analysed fungal species. In contrast, a protein named U4/U6–20 kDa (Cph1p) is found to have a paralog in some species including *B. botryobasidium*, *U. maydis* and *R. irregularis* (Supplementary Table 2).

Pfam domain analyses of these proteins have identified similar domains for each protein in the species used in this study (Figure 3 *a*). The PRPF3 protein from this group carries two domains: PRP3 (PF08572) and DUF1115 (PF06544). The PRP3 domain is a U4/U6-associated splicing factor[43], whereas the functions of DUF1115 are still unknown. The PPIH protein consists of a single domain named proisomerase (cyclophilin type peptidyl-prolyl *cis-trans* isomerase, PF00160), which facilitates chaperone and cell signalling[44]. The NHP2L1 protein has aribosomalL7Ae (PF01248) domain, whereas the PRPF4 protein includes a PRP4 (PF08799) and a WD40 (PF00400) domain. PRP4 is a U4/U6 snRNP that is involved in pre-mRNA processing, whereas WD40 is involved in several functions[45]. The PRPF31 protein includes three domains named NOSIC (PF08060), Nop (PF01798) and Prp31 (PF09785). The Nop domain is a part of pre-RNA processing ribonucleoproteins, whereas Prp31 is required for the U4/U6.U5 tri-snRNP formation[46].

In the U4/U6.U5 tri-snRNP group, all human proteins have orthologs in the analysed fungal species except the tri-snRNP 65 kDa (Sad1p) protein. No ortholog for this protein was identified in *N. crassa* and *S. cerevisiae*.

Moreover, the tri-snRNP 110 kDa (Snu66p) geneis duplicated in *R. solani* (Supplementary Table 3).

In the different proteins of this group, similar domains were identified for each human and corresponding fungal protein (Figure 3 *b*). The SART-1 protein contains a domain (PF03343), which has been reported to be involved in cell cycle arrest and pre-mRNA splicing[47]. The SNRNP27 protein has a domain of unknown function. The USP39 protein carries two domains, zf-UBP (PF02148) and UCH (PF00443). The Zf-UBP domain is known to cleave isopeptide bonds between ubiquitin moieties[48,49].

Fungal non-snRNP proteins related to spliceosomal assembly and splicing

We queried 105 human non-snRNP proteins and searched for homologs in different fungal genomes. With non-snRNP spliceosomal proteins, several human proteins have homologs in the analysed fungal species, except in *S. cerevisiae*. There was quite a significant difference in the number of non-snRNPs identified in filamentous fungi and yeast. All non-snRNP proteins were compiled in different tables (Supplementary Tables 4–10).

Three SR and SR related proteins were identified in all filamentous fungi. However, no homologs of these three proteins were identified in yeast (Supplementary Table 4). Pfam domain analysis depicted the dominant presence of a RRM domain in all three proteins (Supplementary Table 11).

Subsequently, seven human PRP19 complex associated proteins were queried on the studied fungal species. All seven proteins have orthologs in the analysed fungal species (Supplementary Table 5). The conservation of this group of proteins is high in all fungal species. Pfam domain analyses identified similar domains for each protein of this group. In the tPrp19 protein, two Pfam domains were identified as Prp19 (PF08606) and WD40 (PF00400). Prp19 forms an oligomer that is essential for spliceosomal assembly[50], whereas a WD40 Pfam domain is found in Prp46 protein. The CDC5 protein comprises two domains namely MybCef (PF11831) and MybDNAbind (PF13921). MybCef is found in Myb-related Cdc5p/Cef1 proteins and plays a significant role in pre-mRNA splicing factor complex[51], whereas MybDNAbind is a DNA binding domain[14]. In the fSap33 protein the ISY (PF06246) domain was identified in all the analysed fungal species. Isy1 is crucial for optimization of splicing[52]. Likewise, in SYF1 and Crn protein, TPR and HAT domains were identified in all the analysed genomes (Supplementary Table 12).

Additionally, we examined status of the catalytic step II and late-acting proteins (Supplementary Table 13), exon junction complex (EJC) proteins (Supplementary Table 14) and other classes of non-snRNPs (Supplementary Table 15) as described in the Supplementary Sections 3–5.

Similarly we have supplemented detailed information about hnRNP proteins (Supplementary Section 6) and about the fungal proteins involved in alternative splicing (Supplementary Section 7). Majority of these proteins are present in fungi with exception in yeasts.

Discussion

In the current study, we have analysed 192 different types of spliceosomal proteins from 8 fungal species belonging to various order divisions like Ascomycetes, Basidiomycetes and Glomeromycota.

There are 49 snRNP proteins which are further classified into different sub-groups like Sm/LSm core proteins, U1, U2, U5, U4/U6, U4/U6.U5 tri-snRNP. In this study, we demonstrated that snRNP proteins are present in fungal species. The extensive presence of snRNPs suggests that snRNPs play a significant role in regulating and catalysing the fundamental mechanism of splicing. However, the number of snRNPs identified in yeast differs slightly from earlier data[8], this only identified only 43 snRNPs. The two additional proteins identified in our study are U5 snRNP specific U5–52 kDa (CD2BP2) and U1 snRNP specific U1A (SNRPA). This may be due to the use of human protein sequences as query, when compared to Yu *et al.*[8], where amoeboid protozoan *Dictyostelium discoideum* protein sequences were used.

We analysed a total of 105 non-snRNP proteins associated with spliceosomal assembly and splicing in different species of fungal classes. The number of non-snRNP orthologs identified in all the analysed filamentous fungi is significantly different when compared to yeast. Several non-snRNP proteins do not have orthologs in yeast.

SR (serine/arginine-rich proteins) and SR-related proteins play a significant role in regulating both alternative splicing and constitutive splicing[53]. All three queried proteins identified orthologs in the selected species. No protein in this group had an ortholog in yeast. This fits the fact that splicing in yeast is a rarely occurring mechanism due to its intron-poor genome[54].

EJC assembly is a significant step of splicing[55]. All five human EJC proteins have orthologs in the analysed fungal species except yeast (Supplementary Section 4). Only one ortholog was identified in yeast. This again may indicate that processing of RNA is more complex in filamentous fungi than in yeast.

A number of other spliceosomal proteins were identified in the examined fungal species. Human proteins with a DExD/H motif have orthologs in all the analysed fungal species except some proteins like DICE (INTS6) and FLJ41215 (DDX26B). Cyclophilins belong to the proteins catalysing peptidyl-prolyl *cis-trans* isomerase activity[44]. Human cyclophilins and polyadenylation motif-containing proteins have orthologs in the fungal species with some exceptions to yeast. This includes CyP60 (PPIL2), CyPJ (PPIL2) and NY-CO-10.

Overall, we found that spliceosomal proteins have different protein domains (Figures 2–3). Majority of these proteins were conserved in several fungal species and most of these proteins possessed more than one protein domain (Figures 2 and 3). Homologous proteins have conserved domains such as in SNRPB2 and SF3B4 (Figure 2) and in PRPF40A/B (Figure 2). The conserved domains of different spliceosomal proteins assisted the orthology assignments for putative homologs in analysed fungal genomes.

Intron-poor organisms lack several of these proteins as exemplified by the dataset from yeast. A similar case was reported for intron-poor *Giardia lamblia*, a human intestinal parasite. Bordonné *et al.*[56] reported on roles of the protein domain of PRP4 in yeast. The yeast share conserved domains of spliceosomal proteins (which are present in yeast) with filamentous fungi, which are conserved across other eukaryotic lineages.

Protein domain analyses of uncharacterized proteins help in annotation, which lead to several benefits for protein annotators and also for understanding of protein functionality. Understanding the domains of these proteins help in detection of these proteins in newer fungal genomes and in other organisms. Moreover, study of conserved domains is a good approach for constructing structural analyses of these proteins. The complete vignette of the spliceosomal mechanism in three dimension space is possible only when we understand of the structural and interacting components of spliceosomes. Recently, a major breakthrough was achieved with the availability of the crystal structure of PRP8, which provides the structural architectures of the spliceosome active site[57]. Similarly, the crystal structure of human U1 snRNP also provided mechanisms of 5′-ss recognition[58]. These are major steps in assembling complete spliceosomal mechanisms, and other computational studies may facilitate in the future direction of research[59]. Similarly, our domain analyses provides a first glimpse that fungal spliceosomal proteins are by and large similar to human counterparts and can play an instrumental role in assembling this mechanism similar to their human homologs.

To the best of our knowledge, this is the first comprehensive and comparative report on spliceosomal proteins and related factors using multiple fungal species. We believe these results will be beneficial for further experiments to analyse the splicing mechanism in fungi.

1. Brow, D. A., Allosteric cascade of spliceosome activation. *Annu. Rev. Genet.*, 2002, **36**, 333–360.
2. Irimia, M. and Roy, S. W., Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol*., 2014, **6**.
3. Calarco, J. A., Zhen, M. and Blencowe, B. J., Networking in a global world: establishing functional connections between neural splicing regulators and their target transcripts. *RNA*, 2011, **17**, 775–791.
4. Hoskins, A. A. and Moore, M. J., The spliceosome: a flexible, reversible macromolecular machine. *Trends Biochem. Sci*., 2012, **37**, 179–188.
5. Ramani, A. K. *et al.*, Genome-wide analysis of alternative splicing in *Caenorhabditis elegans*. *Genome Res*., 2011, **21**, 342–348.
6. Will, C. L. and Luhrmann, R., Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol*., 2011, **3**.
7. Will, C. L. and Lührmann, R., *The RNA World*, Cold Spring Harbor Laboratory Press, New York, 2006.
8. Yu, B. *et al.*, Spliceosomal genes in the *D. discoideum* genome: A comparison with those in *H. sapiens*, *D. melanogaster*, *A. thaliana* and *S. cerevisiae*. *Protein Cell*, 2011, **2**, 395–409.
9. Beggs, J. D., Lsm proteins and RNA processing. *Biochem. Soc. Trans*., 2005, **33**, 433–438.
10. He, W. and Parker, R., Functions of lsm proteins in mRNA degradation and splicing. *Curr. Opin. Cell Biol*., 2000, **12**, 346–350.
11. Salgado-Garrido, J., Bragado-Nilsson, E., Kandels-Lewis, S. and Seraphin, B., Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. *EMBO. J*., 1999, **18**, 3451–3462.
12. Fromont-Racine, M., Rain, J. C. and Legrain, P., Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet*., 1997, **16**, 277–282.
13. Klein Gunnewiek, J. M., Hussein, R. I., van Aarssen, Y., Palacios, D., de Jong, R., van Venrooij, W. J. and Gunderson, S. I., Fourteen residues of the U1 snRNP-specific U1a protein are required for homodimerization, cooperative RNA binding, and inhibition of polyadenylation. *Mol. Cell Biol*., 2000, **20**, 2209–2217.
14. Aasland, R., Stewart, A. F. and Gibson, T., The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-CoR and TFIIIB. *Trends Biochem. Sci*., 1996, **21**, 87–88.
15. Chaudhury, A., Chander, P. and Howe, P. H., Heterogeneous nuclear ribonucleoproteins (hnrNPS) in cellular processes: focus on hnRNP E1's multifunctional regulatory roles. *RNA*, 2010, **16**, 1449–1462.
16. Dreyfuss, G., Matunis, M. J., Pinol-Roma, S. and Burd, C. G., hnRNP proteins and the biogenesis of mRNA. *Annu. Rev. Biochem*., 1993, **62**, 289–321.
17. Mourelatos, Z., Abel, L., Yong, J., Kataoka, N. and Dreyfuss, G., Smn interacts with a novel family of hnRNP and spliceosomal proteins. *Embo. J*., 2001, **20**, 5443–5452.
18. Miau, L. H., Chang, C. J., Shen, B. J., Tsai, W. H. and Lee, S. C., Identification of heterogeneous nuclear ribonucleoprotein K (hnRNP K) as a repressor of C/EBPβ-mediated gene activation. *J. Biol. Chem*., 1998, **273**, 10784–10791.
19. Mayeda, A. and Krainer, A. R., Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. *Cell*, 1992, **68**, 365–375.
20. Finn, R. D., Clements, J. and Eddy, S. R., HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*., 2011, **39**, W29–W37.
21. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S., MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol*., 2011, **28**, 2731–2739.
22. Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K., Li, J. and Nagai, K., Crystal structure of human spliceosomal U1 snRNP at 5.5 a resolution. *Nature*, 2009, **458**, 475–480.
23. Forch, P., Puig, O., Martinez, C., Seraphin, B. and Valcarcel, J., The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5′ splice sites. *EMBO J*., 2002, **21**, 6882–6892.
24. Query, C. C., Bentley, R. C. and Keene, J. D., A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. *Cell*, 1989, **57**, 89–101.
25. Kobe, B. and Kajava, A. V., The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol*., 2001, **11**, 725–732.
26. Kramer, A., Mulhauser, F., Wersig, C., Groning, K. and Bilbe, G., Mammalian splicing factor SF3a120 represents a new member of

the SURP family of proteins and is homologous to the essential splicing factor Prp21p of *Saccharomyces cerevisiae*. *RNA*, 1995, **1**, 260–272.
27. Kuwasako, K., He, F., Inoue, M., Tanaka, A., Sugano, S., Guntert, P., Muto, Y. and Yokoyama, S., Solution structures of the SURP domains and the subunit-assembly mechanism within the splicing factor SF3a complex in 17S U2 snRNP. *Structure*, 2006, **14**, 1677–1689.
28. Will, C. L., Urlaub, H., Achsel, T., Gentzel, M., Wilm, M. and Luhrmann, R., Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b dead-box protein. *EMBO. J*., 2002, **21**, 4978–4988.
29. Trappe, R., Ahmed, M., Glaser, B., Vogel, C., Tascou, S., Burfeind, P. and Engel, W., Identification and characterization of a novel murine multigene family containing a PHD-finger-like motif. *Biochem. Biophys. Res. Commun.*, 2002, **293**, 816–826.
30. Hryciw, T., Tang, M., Fontanie, T. and Xiao, W., MMS1 protects against replication-dependent DNA damage in saccharomyces cerevisiae. *Mol. Genet. Genomics*, 2002, **266**, 848–857.
31. Li, Y., Chen, Z. Y., Wang, W., Baker, C. C. and Krug, R. M., The 3′-end-processing factor CPSF is required for the splicing of single-intron pre-mRNAS *in vivo*. *RNA*, 2001, **7**, 920–931.
32. Freund, C., Dotsch, V., Nishizawa, K., Reinherz, E. L. and Wagner, G., The GYF domain is a novel structural fold that is involved in lymphoid signaling through proline-rich sequences. *Nat. Struct. Biol.*, 1999, **6**, 656–660.
33. Nishizawa, K., Freund, C., Li, J., Wagner, G. and Reinherz, E. L., Identification of a proline-binding motif regulating CD2-triggered T lymphocyte activation. *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 14897–14902.
34. Reuter, K., Nottrott, S., Fabrizio, P., Luhrmann, R. and Ficner, R., Identification, characterization and crystal structure analysis of the human spliceosomal U5 snRNP-specific 15 KD protein. *J. Mol. Biol*., 1999, **294**, 515–525.
35. Jermy, A. J., Willer, M., Davis, E., Wilkinson, B. M. and Stirling, C. J., The BRL domain in Sec63p is required for assembly of functional endoplasmic reticulum translocons. *J. Biol. Chem*., 2006, **281**, 7899–7906.
36. Ponting, C. P., Proteins of the endoplasmic-reticulum-associated degradation pathway: Domain detection and function prediction. *Biochem. J*., 2000, **351(**Pt 2), 527–535.
37. Aubourg, S., Kreis, M. and Lecharny, A., The dead box RNA helicase family in *Arabidopsis thaliana*. *Nucleic Acids Res*., 1999, **27**, 628–636.
38. de la Cruz, J., Kressler, D. and Linder, P., Unwinding RNA in *Saccharomyces cerevisiae*: dead-box proteins and related families. *Trends Biochem. Sci*., 1999, **24**, 192–198.
39. Urushiyama, S., Tani, T. and Ohshima, Y., Isolation of novel pre-mRNA splicing mutants of *Schizosaccharomyces pombe*. *Mol. Gen. Genet*, 1996, **253**, 118–127.
40. Blatch, G. L. and Lassle, M., The tetratricopeptide repeat: a structural motif mediating protein–protein interactions. *Bioessays*, 1999, **21**, 932–939.
41. Staub, E., Fiziev, P., Rosenthal, A. and Hinzmann, B., Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire. *Bioessays*, 2004, **26**, 567–581.
42. Ritchie, D. B., Schellenberg, M. J., Gesner, E. M., Raithatha, S. A., Stuart, D. T. and Macmillan, A. M., Structural elucidation of a Prp8 core domain from the heart of the spliceosome. *Nat. Struct. Mol. Biol*., 2008, **15**, 1199–1205.
43. Ayadi, L., Callebaut, I., Saguez, C., Villa, T., Mornon, J. P. and Banroques, J., Functional and structural characterization of the Prp3 binding domain of the yeast Prp4 splicing factor. *J. Mol. Biol.*, 1998, **284**, 673–687.
44. Wang, P. and Heitman, J., The cyclophilins. *Genome Biol*., 2005, **6**, 226.
45. Stirnimann, C. U., Petsalaki, E., Russell, R. B. and Muller, C. W., WD40 proteins propel cellular networks. *Trends Biochem. Sci*., 2010, **35**, 565–574.
46. Weidenhammer, E. M., Singh, M., Ruiz-Noriega, M. and Woolford Jr., J. L., The Prp31 gene encodes a novel protein required for pre-mRNA splicing in *Saccharomyces cerevisiae*. *Nucleic Acids Res*., 1996, **24**, 1164–1170.
47. Wilkinson, C. R., Dittmar, G. A., Ohi, M. D., Uetz, P., Jones, N. and Finley, D., Ubiquitin-like protein HUB1 is required for pre-mRNA splicing and localization of an essential splicing factor in fission yeast. *Curr. Biol*., 2004, **14**, 2283–2288.
48. Hershko, A. and Ciechanover, A., The ubiquitin system. *Annu. Rev. Biochem*., 1998, **67**, 425–479.
49. Wilkinson, K. D., Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. *FASEB J*., 1997, **11**, 1245–1256.
50. Grillari, J. *et al.*, SNEV is an evolutionarily conserved splicing factor whose oligomerization is necessary for spliceosome assembly. *Nucleic Acids Res*., 2005, **33**, 6868–6883.
51. Ohi, M. D., Link, A. J., Ren, L., Jennings, J. L., McDonald, W. H. and Gould, K. L., Proteomics analysis reveals stable multiprotein complexes in both fission and budding yeasts containing Myb-related Cdc5p/Cef1p, novel pre-mRNA splicing factors, and snRNAS. *Mol. Cell Biol*., 2002, **22**, 2011–2024.
52. Dix, I., Russell, C., Yehuda, S.B., Kupiec, M. and Beggs, J. D., The identification and characterization of a novel splicing protein, Isy1p of *Saccharomyces cerevisiae*. *RNA*, 1999, **5**, 360–368.
53. Shepard, P. J. and Hertel, K. J., The SR protein family. *Genome Biol*., 2009, **10**, 242.
54. Kempken, F., Alternative splicing in ascomycetes. *Appl. Microbiol. Biotechnol.*, 2013, **97**, 4235–4241.
55. Sauliere, J., Haque, N., Harms, S., Barbosa, I., Blanchette, M. and Le Hir, H., The exon junction complex differentially marks spliced junctions. *Nat. Struct. Mol. Biol*., 2010, **17**, 1269–1271.
56. Bordonné, R., Banroques, J., Abelson, J. and Guthrie, C., Domains of yeast U4 spliceosomal RNA required for Prp4 protein binding, snRNP–snRNP interactions, and pre-mRNA splicing *in vivo*. *Genes Dev.*, 1990, **4**, 1185–1196.
57. Galej, W. P., Oubridge, C., Newman, A. J. and Nagai, K., Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature*, 2013, **493**, 638-+.
58. Kondo, Y., Oubridge, C., van Roon, A. M. M. and Nagai, K., Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5′ splice site recognition. *Elife*, 2015, **4**.
59. Korneta, I., Magnus, M. and Bujnicki, J. M., Structural bioinformatics of the human spliceosomal proteome. *Nucleic Acids Res*., 2012, **40**, 7046–7065.

ACKNOWLEDGEMENT. We thank Julia Friman for editing this manuscript.

Received 12 August 2017; revised accepted 20 November 2017

doi: 10.18520/cs/v114/i08/1677-1686