

Pre-mRNA Splicing: Awash in a Sea of Proteins

Review

Melissa S. Jurica and Melissa J. Moore*

Howard Hughes Medical Institute
Department of Biochemistry
Brandeis University
415 South Street
Waltham, Massachusetts 02454

What's in a spliceosome? More than we ever imagined, according to recent reports employing proteomics techniques to analyze this multi-megadalton machine. As of 1999, around 100 splicing factors were identified (Burge et al., 1999); however, that number has now nearly doubled due primarily to improved purification of spliceosomes coupled with advances in mass spectrometry analyses of complex mixtures. Gratifyingly, most of the previously identified splicing factors were found in the recent mass spec studies. Nonetheless, the number of new proteins emerging with no prior connection to splicing was surprising. Without functional validation, it would be premature to label these proteins as bona fide splicing factors. Yet many were identified multiple times in complexes purified under diverse conditions or from different organisms. Another recurring theme regards the dynamic nature of spliceosomal complexes, which may be even more intricate than previously thought.

A Tsunami of Purified Spliceosomes

The spliceosome is a collection of snRNAs and proteins recruited to nascent transcripts (pre-mRNAs) to carry out intron excision. Until last year, the prevailing dogma in the splicing field posited that as each new intron emerges from the transcription machinery, its recognition and removal requires complete reassembly of a new spliceosome from component parts. These parts include five U snRNPs (U1, U2, U4, U5, and U6), each containing a small stable RNA bound by several proteins, plus numerous other less stably-associated splicing factors. Within the assembled spliceosome, intron excision occurs in two chemical steps: first 5' splice site cleavage and lariat formation, then 3' splice site cleavage and exon ligation. Following exon ligation, complete spliceosome disassembly would then free up its components for de novo synthesis of other spliceosomes. This so-called "splicing cycle" (Figure 1), which appears in almost every modern molecular biology textbook, was mostly derived from in vitro studies wherein splicing is uncoupled from other pre-mRNA processing events. In such experiments, spliceosome assembly proceeds through a series of short-lived intermediate stages dubbed E/CC, A, B, and C. These subcomplexes can be distinguished by their different mobilities in native gels or density gradients, their snRNP composition, and the chemical state of the pre-mRNA (i.e., whether or not either chemical step of splicing has occurred).

Because of the difficulties inherent in stabilizing and

purifying any one spliceosomal subcomplex in vitro, the most common proteomics approach to date has been to work with mixtures of assembly intermediates (Neubauer et al., 1998; Rappsilber et al., 2002; Zhou et al., 2002). The goal here is to enumerate every protein involved no matter at what stage. Two of these studies employed pre-mRNAs containing randomly incorporated biotins as affinity tags (Neubauer et al., 1998; Rappsilber et al., 2002). However, the near-covalent nature of the biotin-avidin interaction necessitates protein elution by denaturation, precluding further structural or functional characterization of intact complexes. More recently, the advent of pre-mRNAs containing aptamer sequences (e.g., tobramycin) or binding sites for an affinity-tagged RNA binding protein (e.g., viral MS2 protein fused to maltose binding protein; MS2-MBP) has made it possible to elute splicing complexes under nondenaturing conditions (Das et al., 2000; Wang and Rando, 1995), and this was the approach taken by Zhou et al. (2002).

Last year also saw purification and mass spec analysis of three individual subcomplexes, each captured in a different way (Hartmuth et al., 2002; Jurica et al., 2002; Makarov et al., 2002). Complex A (a.k.a. the pre-spliceosome) is an early assembly intermediate. It contains primarily U1 and U2 snRNPs bound to unspliced pre-mRNA and was assembled on a pre-mRNA substrate immobilized on a solid support via a tobramycin aptamer tag. For unknown reasons, this solid-state assembly arrested spliceosome progression at A complex (Hartmuth et al., 2002). Complex B* is the form of the spliceosome poised to perform the first chemical step. In addition to unspliced pre-mRNA, it contains U2, U5, and U6 snRNAs. This complex was isolated by immunoaffinity against SKIP, a protein that joins the spliceosome concomitant with the U4/U6-U5 triple-snRNP. Interestingly, anti-SKIP antibodies also pull down a 35S complex containing U5 snRNA and a distinct set of proteins that includes splicing factors associated with later stages of spliceosome assembly, but not the pre-mRNA substrate. Since the 35S U5 complex could also be isolated from splicing reactions lacking any pre-mRNA substrate, it may represent a spliceosome disassembly intermediate (Makarov et al., 2002). Like complex B*, complex C contains U2, U5, and U6 snRNAs, but represents a stage subsequent to the first chemical step (i.e., it contains splicing intermediates). It was accumulated on a pre-mRNA incapable of completing exon ligation and purified via MS2-MBP bound to intron (Jurica et al., 2002).

Given the dynamic nature of splicing complexes in vitro, purification of endogenous higher-order complexes at first appeared unlikely. Two papers published last year, however, are requiring us to reevaluate the extent to which the in vitro-derived splicing cycle reflects the actual goings on in vivo. First, Abelson and coworkers (Stevens et al., 2002) reported purification of a pentasnRNP complex from *S. cerevisiae* that may represent a preassembled spliceosome. The implication is that intron removal in vivo may be orchestrated by preex-

*Correspondence: mmoore@brandeis.edu

discern which are truly spliceosomal and which are merely contaminants. In mass spec experiments, some proteins are almost always detected because of their sheer abundance in the cell. Whereas some labs included ribosomal proteins, EF-Tu, tubulin, and/or heat shock proteins in their lists of spliceosomal proteins, others may have also detected these species but ruled them out as a matter of course because they are typical mass spec contaminants. Similarly, because splicing complexes necessarily contain RNA, and RNA binding proteins abound in cells, many polypeptides will copurify with spliceosomes simply because they bind RNA. One strategy for weeding out such proteins was to perform parallel purifications using a tagged pre-mRNA, but under conditions that either do or do not support splicing (Jurica et al., 2002; Zhou et al., 2002). This resulted in over 30 proteins, including the hnRNP proteins, being classified as associating with RNA independent of splicing. A similar strategy applied to the Pol II preinitiation complex showed that only 15% of the 326 proteins initially identified as being preinitiation complex associated are specifically recruited upon TATA binding protein (TBP) addition (Ranish et al., 2003). Of course repetition also helps. Perhaps the most conservative approach for defining spliceosomal proteins was taken by the Gould laboratory (Ohi et al., 2002). They affinity-tagged six different proteins, including some of the newly identified ones, and limited their final list to polypeptides that appeared in most or all purifications.

Another issue concerns how to distinguish sequence-specific pre-mRNA splicing factors from core components required for splicing of every intron. Since mammalian splice site consensus sequences are rather degenerate, recognition of individual sites often depends on their sequence context. Many splice sites are flanked by exonic and intronic splicing enhancer elements, which bind an array of auxiliary factors that assist in recruiting the core splicing machinery. Thus, spliceosomes assembled on different pre-mRNA substrates are likely to contain a different complement of these auxiliary factors. The proteins that define the catalytic core, however, should be present regardless of pre-mRNA sequence. To date, this issue remains largely unaddressed. Only a limited number of pre-mRNA substrates work well for splicing *in vitro*, with derivatives of Adenovirus Major Late (AdML) intron 1 being perhaps the most widely studied. Thus all groups assembling mammalian spliceosomes *in vitro* employed some AdML variant. However, Rappsilber et al. (2002) and Zhou et al. (2002) also analyzed complexes assembled on a second substrate. Whereas Zhou et al. reported only those proteins present in both samples as a means to weed out sequence-specific proteins, Rappsilber et al. combined their results to identify as many potential splicing factors as possible. Obviously, many more such analyses will need to be performed before we can definitively assign new proteins as sequence-specific or core.

In addition to proteins recognizing *cis*-acting regulatory sequences, other auxiliary factors provide crucial communication links between the splicing machinery and other processes such as transcription, capping, and 3'-end formation (Proudfoot et al., 2002). One way to differentiate such auxiliary factors from catalytic core components is that the latter should maintain their asso-

ciation even under stringent purification conditions. While some groups used very mild conditions in hopes of purifying holospliceosomes (Rappsilber et al., 2002; Stevens et al., 2002; Zhou et al., 2002), others treated their complexes to high salt washes, EDTA, and/or heparin in an attempt to identify just the core components (Hartmuth et al., 2002; Jurica et al., 2002; Makarov et al., 2002). Another characteristic of core components is that they should be present in stoichiometric amounts. Therefore, it would be very informative to know the relative abundance of each protein detected. Proteins are generally identified in mass spec by their unique tryptic peptides, and several groups reported the number of unique peptides detected for each protein. Yet, while one might assume that more peptides mean more protein, this is not necessarily the case. Unfortunately, without internal standards, mass spec is not particularly quantitative. Larger proteins generally yield more detectable peptides than smaller ones simply because they contain a greater number of peptides. So although finding multiple peptides from a protein increases confidence in its presence, it is inadvisable to make any strong conclusions about relative protein abundance solely from the number of peptides identified. Currently, relative staining intensities on denaturing gels or quantitative Western blotting remain the most reliable way to estimate relative protein abundances. It is also possible to very precisely determine protein abundance by spiking mass spec samples with isotopically labeled peptide standards (Gerber et al., 2003), but this technique has yet to be applied on a large scale.

Core versus auxiliary associations aside, for new proteins coming out of the mass spec screens, one is still faced with the task of resolving the bona fide splicing factor versus contaminant issue. Of course, functional and genetic data provide the most compelling evidence for a role in splicing, but given the large number of new proteins, such evidence may not be immediately forthcoming. Furthermore, many of the newly identified proteins have no clear homologs in *S. cerevisiae* or *S. pombe*, the most genetically malleable organisms. Almost certainly the laboratories reporting these new proteins are busy performing genetic knock-down and/or immunodepletion experiments, but in the meantime what can one say? This is, what information can be gleaned from examination of the mass spec data alone? Although caveats abound for interpretation of the specific proteins found in any one experiment, what proves more useful is to look for patterns across the entire data set. For example, if a new protein appears in multiple complexes purified under different conditions and/or from different organisms, one begins to feel comfortable that it is unlikely a spurious contaminant. Tables 1–3 provide such a comparison for the proteins so far identified in multiple mass spec analyses of purified splicing complexes. Table 1 contains the cloned splicing factors included in the 1999 review by Burge et al. These proteins are organized primarily by their U snRNP associations. Table 2 lists additional polypeptides reported by more than one group in the recent mass spec experiments outlined above. A number of these proteins have demonstrated roles in splicing, while others are involved in pre-mRNA processing events linked to splicing. Still, the majority of these proteins are novel with respect to

Table 1. Previously Known Spliceosomal Proteins

Human Name	Ensemble Accession	<i>S. cerevisiae</i> Name																			
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
snRNA																					
U1 snRNA			na	na	na	A				P					na	na	na	na	na	na	
U2 snRNA			na	na	na	A	B*	C	Cl	P		12			na	na	na	na	na	na	
U4 snRNA			na	na	na					P			Tr		na	na	na	na	na	na	
U5 snRNA			na	na	na		B*	C	Cl	P	35		Tr	H	na	na	na	na	na	na	
U6 snRNA			na	na	na		B*	C	Cl	P			Tr		na	na	na	na	na	na	
Core snRNP proteins																					
SmB/B'	125835	Smb1	S	S	S	A	B*	C	Cl	P	35	12	Tr	H	T1			T4	T5	T6	
SmD1	167088	Smd1		S	S	A	B*	C	Cl	P	35	12	Tr			T2		T4	T5	T6	
SmD2	125743	Smd2		S	S	A	B*	C	Cl	P	35	12	Tr		T1	T2	T3	T4	T5	T6	
SmD3	100028	Smd3		S	S	A	B*	C	Cl	P	35	12	Tr	h		T2		T4	T5	T6	
SmE1	176773	Sme1		S	S	A	B*	C		P	35	12	Tr								
SmF1	139343	Smf1		S	S	A	B*	C	Cl	P	35	12	Tr						T5		
SmG1	143977	Smx2		S	S	A	B*	C	Cl	P	35	12	Tr								
LSM2	111987	Lsm2		S	S		B*	C		P			Tr		T1					T6	
LSM3	170860	Lsm3		S	S		B*	C		P					T1						
LSM4	130520	Lsm4		S	S					P			Tr		T1			T4	T5	T6	
LSM5	106355	Lsm5											Tr		T1					T6	
LSM6	164167	Lsm6		S	S					P			Tr		T1					T6	
LSM7	130332	Lsm7		S	S					P			Tr		T1						
LSM8	128534	Lsm8		S						P			Tr								
U1 snRNP specific proteins																					
U1-70kD	104852	Snp1	S	S	S	A				P						T2		T4	T5	T6	
U1-A	077312	Mud1	S	S	S	A				P				h		T2		T4	T5	T6	
U1-C	124562	Yhc1		S	S	A				P								T4	T5	T6	
FBP11	123596	Prp39								P						T2		T4	T5	T6	
		Prp40			S					P						T2		T4	T5	T6	
		Snu56									P					T2		T4	T5	T6	
		Nam8									P					T2			T5		
		Snu71									P					T2		T4	T5	T6	
Snu65									P					T2		T4	T5	T6			
U2 snRNP specific proteins																					
U2-A'	131876	Lea1	S	S	S	A	B*	C	Cl	P		12		h			T3	T4	T5	T6	
U2-B''	125870	Msl1	S	S	S	A	B*	C	Cl	P		12						T4	T5	T6	
SF3a60	nIm	Prp9	S	S	S	A	B*	C	Cl	P		12		h				T4	T5	T6	
SF3a66	104897	Prp11	S	S	S	A	B*	C	Cl	P		12						T4	T5	T6	
SF3a120	099995	Prp21		S	S	A	B*	C	Cl	P		12		H			T3	T4	T5	T6	
SF3b49	143368	Hsh49	S	S	S	A	B*	C	Cl	P		12						T4			
SF3b145	087365	Cus1	S	S	S	A	B*	C	Cl	P		12		H				T4	T5	T6	
SF3b130	nIm	Rse1	S	S	S	A	B*	C	Cl	P		12		H		T2	T3	T4	T5	T6	
SF3b155	115524	Hsh155		S	S	A	B*	C	Cl	P		12		H			T3	T4	T5	T6	
p14	115128	Snu17		S	S	A	B*			P		12									
U5 snRNP specific proteins																					
PRP8	174231	Prp8		S	S	A	B*	C	Cl	P	35		Tr	h	(T1)	(T2)	(T3)	(T4)	(T5)	(T6)	
U5-200kD	144028	Brr2		S	S	A	B*	C	Cl	P	35		Tr	H							
U5-116kD	108883	Snu114		S	S	A	B*	C	Cl	P	35		Tr	H	T1		T3	T4	T5	T6	
U5-102kD	101161	Prp6		S	S	A	B*	C		P			Tr	h	T1			T4	T5	T6	
U5-100kD	174243	Prp28		S	S	S	A			C				h						T6	
U5-40kD	060688		S	S	S		B*	C	Cl		35			h							
U5-15kD	141759	Dib1			S					P			Tr		T1			T4	T5	T6	
U4/U6 snRNP specific proteins																					
HPRP3	117360	Prp3	S	S	S					P			Tr					T4	T5	T6	
HPRP4	136875	Prp4	S	S	S			C		P			Tr		T1			T4	T5	T6	
RY-1	124380				S																
USA-Cyp	171960	Cpr1			S																
15.5 tri-snRNP	100138	Snu13		S	S					P			Tr								
Miscellaneous splicing factors																					
U2AF65	063244	Mud2	S	S	S	A						12		H							
SF1	168066	Msl5	S	S	S																
CBP20	114503	Cbc2		S	S	A	B*	C								T2			T5		
CBP80	136937	Sto1		S	S	A	B*	C						H		T2	T3	T4	T5	T6	

(continued)

Table 1. Continued

Human Name	Ensemble Accession	<i>S. cerevisiae</i> Name	<i>S. cerevisiae</i>																	
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
U2AF35	160201		S	S	S	A						12								
ASF/SF2	136450			S	S	A		C				12								
UAP56	173539	Sub2			S										T2		T4	T5	T6	
PRP5	145833	Prp5		S	S	A						12								
Tat-SF1	102241	Cus2			S						P									
PTB	011304			S										H						
PRP19	110107	Prp19	S	S	S	A		C	Cl	P	35			H		T3	T4	T5	T6	
PRP31	105618	Prp31		S	S					P			Tr		T1		T4	T5	T6	
		Snt309								Cl	P						T3	T4	T5	T6
DDX16	137333	Prp2		S	S			C								T3				
PRP16	140829	Prp16		S	S			C												
PRP17	168438	Prp17		S	S			C	Cl		35					T3				
SLU7	164609	Slu7		S	S			C	Cl											T6
PRP18	165630	Prp18																		T6
PRP22	067596	Prp22		S	S		B*	C	Cl							T3				
EWS	n1m			S				C												
		Prp38								P			Tr		T1					T6
PRP43	109606	Prp43		S	S	A	B*	C				12		H	T1		T3	T4	T5	T6
PRP24	075856	Prp24													T1					T6
DDX3	124487	Ded1			S					P										
		Npl3								P						T2			T5	

Derived from Burge et al. (1999). The Ensemble accession number designates the human locus for the protein and is preceded in all cases by ENSG00000 (Inm: locus not mapped). When known, the *S. cerevisiae* homolog is given. The numbered columns represent proteins found in the following complexes: 1, 2, 3: Mixed spliceosomes (S) assembled in vitro from human nuclear extract; (1: Neubauer et al., 1998); (2: Rappsilber et al., 2002); (3: Zhou et al., 2002). 4, 5, 6: Individual complexes assembled in vitro from human nuclear extract: A complex (A) (4: Hartmuth et al., 2002); B* complex (B*) (5: Makarov et al., 2002); C complex (C) (6: Jurica et al., 2002) (included in column 6 are proteins not previously published identified by four or fewer unique tryptic peptides or due to their presence in both C and H complexes. A small c denotes proteins identified by only a single peptide). 7: Endogenous C complex-like (Cl) U2/U5/U6 complex from *S. cerevisiae* and *S. pombe* (Ohi et al., 2002). 8: Endogenous penta-snRNP (P) from *S. cerevisiae* (Stevens et al., 2002). 9: U5 snRNP 35S complex (35) from human extracts (Makarov et al., 2002). 10: U2 12S complex (12) from human extracts (Will et al., 2002). 11: U4/U6:U5 triple snRNP (Tr) from *S. cerevisiae* (Gottschalk et al., 1999; Stevens and Abelson, 1999). 12: H complex (H) assembled in vitro from human extracts (Jurica et al., 2002). A small h denotes proteins identified by only a single peptide. 13–18: Proposed complexes (T1–T6) containing splicing factors from *S. cerevisiae* based on systematic TAP tagging of 589 proteins followed by purification and mass spec identification of associated proteins (Gavin et al., 2002). T1–T6 correspond to complexes 128, 129, 155, 158, 160, and 161, respectively, from Supplemental Data S3 in that reference.

splicing. Table 3 contains proteins primarily associated with H complex, defined here as the complement of proteins bound to RNA in the absence of splicing.

Appraising the Nuggets

As one scans across the Tables 1 and 2, several themes emerge. One theme is the numerous proteins present in almost every splicing complex. Most comprise what can be thought of as the core splicing machinery—components necessary for splicing of every intron. In Table 1, the most striking of these are the Sm proteins, core components of U1, U2, U4, and U5 snRNPs, and all nine proteins specific to U2 snRNP. Most U5 snRNP proteins were also found in the majority of complexes. Yet, the observation that U5-100 kDa was detected in some complexes and not others has been suggested to reflect a previously unsuspected dynamic rearrangement of this snRNP over the course of the splicing cycle (Makarov et al., 2002). Indeed, since the presence and/or absence of numerous characterized splicing factors follows expectations in many cases (e.g., the presence in C, but not in A or B* complexes, of proteins known to be solely required for the second step of splicing, as well as the absence of U1 proteins from B* and C complexes), it is tempting to extend this interpretation to other proteins. However, one must be careful here. A failure to detect a protein in any single round of mass

spectrometry does not necessarily indicate its absence. For example, many of the core U6 snRNP proteins (Lsm's) were undetectable in complexes that clearly contained U6 snRNA. Whether this reflects a change in Lsm association as splicing proceeds or merely difficulties inherent in mass spec detection of small proteins remains an open question. Further, even if a protein is absent from a particular complex, this does not necessarily imply a lack of function at that stage. Many loosely associated factors are invariably lost during purification. An example of this type is the Prp18 protein. Despite a well-characterized function in the second step of splicing, Prp18 was undetectable in purified C complex either by mass spec or Western blotting (Jurica et al., 2002).

A number of new proteins identified as splicing factors since the Burge et al. review (1999) also appear to be core components (Table 2). For example, a set of proteins that form a stable complex with Prp19 were all well represented in splicing complexes subsequent to A complex, as well as in both the yeast penta-snRNP and 35S U5 snRNP. In *S. cerevisiae*, this so-called Prp19 complex joins the spliceosome concurrent with or just following the disassociation of U4 snRNA (Chen et al., 2002; Ohi and Gould, 2002; Tarn et al., 1994). A similar complex was discovered in mammalian extracts by immunoprecipitation of the CDC5 protein (Ajuh et al., 2000). Most of the proteins in the PRP19 complex have

Table 2. New Spliceosomal Proteins

Human Name	Ensemble Accession	<i>S. cerevisiae</i> Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Proteins associated with the Prp19 complex																				
GDC5	096401	Cef1 (Cdc5)	S		S		B*	C	Cl	P	35			H			T3	T4		
ISY1	172780	Isy1 (Cwf12)		S	S		B*	C	Cl	P	35			h			T3			
SYF1	076924	Syf1 (Cwf3)		S	S		B*	C	Cl	P	35						T3			T6
CRN	101343	Clf1 (Cwf4)	S	S	S		B*	C	Cl	P	35	T1		T3	T4	T5	T6			
GCIP-IP	117614	Syf2 (C3E7.13C)		S	S		B*	C	Cl	P	35									
PRL1	171566	Prp46 (Cwf1)		S	S		B*	C	Cl	P	35			h	T1		T3	T4	T5	T6
BCAS2	116752	(Cwf7)	S	S	S		B*		Cl		35									
		Ntc20							Cl		P									
		Cwc2 (Cwf2)							Cl		P						T3	T4	T5	T6
Proteins with demonstrated roles in splicing																				
SKIP	100603	Prp45 (Cwf13)	S	S	S		B*	C	Cl	P	35			H			T3			
ECM2	086589	Ecm2 (Cwf5)		S	S		B*	C	Cl	P	35			h			T3	T4	T5	T6
SART1	175467	Snu66		S	S	A	B*	C		P			Tr		T1			T4	T5	T6
p68	108654	Dbp2	S	S	S	A	B*	C						H						
SPF45	134453		S	S	S	A					12									
SPF30	119953		S	S	S	A					12									
PSF	116560			S			B*							h						
FLJ31121	146007	Snu23					B*		P				Tr		T1					T6
SAD1	168883	Sad1		S	S				P											
LUC7	007392	Luc7 (Luc7)			S				P							T2		T4	T5	T6
		Spp381							P				Tr							T6
SR proteins																				
SRm300	167978			S	S			C												
SRm160	133226			S	S			c												
SC35	161547			S	S	A														
SRp40	100650				S	A														
SRp55	124193			S	S	A														
SRp75	116350				S			c												
SRp30c	111786			S	S	A	B*													
9G8	115875			S	S	A	B*	c												
SRp54	116754			S	S															
SFRS10	136527				S			c												
SRp20	112081			S	S	A	B*													
Proteins with roles in pre-mRNA metabolism processes linked to splicing																				
REF	141592	Yra1	S	S	S	A	B*	c						h						
RNPS1	167971				S			c						h						
Y14	131795				S			C												
MAGOH	162385				S			c												
hTHO2	125676	Rlr1		S	S															
hHPR1	079134			S	S			C						h						
HsKin17	151657	Rts2p		S			B*													
ASR2B	087087			S	S	A														
KIAA0983	100296			S	S															
C21orf66	159086			S	S															
PAB2	100836		S	S	S		B*	c						H						
CF I-68kD	111605			S	S															
CF I-25kD	167005			S	S															
CPSF 160K	071894			S			B*													
Proteins containing a DEAD/H box helicase motif																				
HDB/DICE1	102786			S			B*													
Abstrakt	146074			S	S		C													
eIF4a3	141543			S	S		B*	C												
DDX35	101452							C			35									
DDX9	135829			S	S	A														
KIAA0052	039123			S	S			C						H						
p72	100201			S	S			C						H						
Proteins with homology to cis-trans prolyl isomerases																				
CypE	084072			S	S		B*	C			35									
KIAA0073	113593			S	S			C												

(continued)

Table 2. Continued

Human Name	Ensemble Accession	<i>S. cerevisiae</i> Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
Cyp60	100023			S	S		B*	C													
PPIL3b	115934			S	S		B*	C													
PPIL1	137168	Cwc27 (Cwf27)		S	S		B*	C	Cl		35										
SDCCAG10	153015			S				C													
Additional proteins novel to splicing																					
KIAA1604	163510	Cwc22 (Cwf22)		S	S			C	Cl								T3			T6	
TIP39	100109	YLR424W		S	S		B*	C						T1			T3	T4	T5	T6	
		Cwc21 (Cwf21)							Cl	P											
G10	106245	Cwc14 (Cwf14)		S	S		B*		Cl												
FLJ10374	105248	Yju2 (Cwf16)							Cl								T3				
MGC13125	137656	Cwc26 (Cwf26)		S	S		B*		Cl												
ZNF183	125352	Cwc24 (Cwf24)						C	Cl												
FLJ10634	104129	Cwc23 (Cwf23)		S					Cl												
SF3b14b	100410	Rds3p		S			A	B*			12										
SPF31	126698		S	S	S	A					12										
CHERP	085872		S	S		A					12										
F23858	105705		S	S		A															
CA150	113649		S	S	S		B*														
SF3b10	169976			S		A	B*				12										
SR140	163714			S	S	A	B*				12										
RBM5	003756			S		A															
E1B-AP5	105323			S		A															
FLJ10805	122692			S	S		B*														
MFAP1	140259			S	S		B*														
KIAA0560	021776	(Cwf11)		S	S		B*	C	Cl		35										
RED protein	113141			S	S		B*	C													
Pinin	100941						B*	C						h							
NOSIP	142546						B*	C													
FLJ10206	076650							C			35										
PUF60	179950			S	S			c			12			h							
DGSI	100056			S	S			C													
Cactin	105298			S	S			C													
FRG1	109536			S				C													
PMSC2	171824			S				C													
RBP 7	076053			S				C									h				
MGC23918	160799	(Cwf18)		S			B*		Cl		35										
SNP70	084463			S	S																
OTT	162775			S	S																
IMP3	136231			S	S																
PRP4 kinase	112739			S	S																
AcinusL	100813			S	S									h							
RNPC2	131051			S	S	A		C						H							
FLJ90157	033030			S				C						H							
NuMA	137497			S				C						H							

S. pombe homologs identified are shown in parentheses following the *S. cerevisiae* name. Numbered columns same as Table 1.

been cloned and shown to be required for splicing (Chen et al., 2002; Ohi and Gould, 2002; Tarn et al., 1994), but at least two have yet to be identified. Perusal of the cumulative mass spec data suggests that any of four proteins, SKIP, ECM2, PPIL1, and KIAA0560, might well be these previously unidentified subunits, as all four exhibit identical distributions to known Prp19 complex components. Although PPIL1 and KIAA0560 have no known functions, both SKIP and ECM2 have already been implicated as splicing factors. Originally characterized as an interacting partner to the Ski oncoprotein and proposed to have a role in transcription (Dahl et al., 1998), SKIP was recently shown to be essential for splicing in *S. cerevisiae* (Albers et al., 2003). Likewise, the Ecm2 protein has been shown to play a role in U4, U6, and U2 snRNA rearrangement during spliceosome activation (Xu and Friesen, 2001).

Another prominent family of core splicing factors consists of DExD/H box proteins, some of which are known RNA helicases. Previous genetic analyses in *S. cerevisiae* had identified eight DExD/H proteins (Prp2p, Prp16p, Prp22p, Prp43p, Brr2p, Prp5p, Prp28p, and Sub2p) (Staley and Guthrie, 1998) as being generally required for pre-mRNA splicing. Each of these proteins has been linked to a specific ATP-requiring step in the in vitro-derived splicing cycle, and all of them were detected in the recent mass spec analyses. Yet, although every known ATP-dependent structural transition in the textbook splicing cycle has already been associated with at least one of these known DExD/H proteins, no fewer than seven more spliceosomal DExD/H proteins have now emerged from the proteomics analyses (Table 2). Interestingly, all of the new proteins were found in mammalian complexes and have no clear homologs in bud-

Table 3. H Complex Proteins

Human Name	Ensemble Accession												
		1	2	3	4	5	6	7	8	9	10	11	12
hnRNP proteins													
hnRNP A1	135486	S	S	H	A		C						H
hnRNPA2/B2	122566	S	S	H	A	B*	C						H
hnRNP A3	176825		S	H	A		C						H
hnRNP C	92199	S	S	H	A	B*	C						H
hnRNP D	138668		S	H									H
hnRNP F	169813		S				C						H
hnRNP G	147274		S			B*	C						H
hnRNP H1	169045		S				C						H
hnRNP K	165119		S	H	A	B*	C						H
hnRNP L	104824		S	H									H
hnRNP M	99783	S	S				C						H
hnRNP R	125944		S	H	A		C		35	12			H
hnRNP U	153187		S		A		C						H
hnRNP RALY	125970			H			C						H
Additional H complex proteins													
Ku70	100419		S										H
PTB	11304		S	H									H
Gry-Rbp	135316		S	H			C						H
DNA binding protein a	60138		S	H			C						H
hUR	66044		S	H	A		C						H
NF45	143621		S	H			C						H
NF90	129351		S	H	A		C						H
MAT3	15479		S	H			C						H
YB-1	65978		S	H	A		C						H
TLS ip	89280		S	H	A		C						H
HSP70-2	126803		S	H			C		35				H
HSP71	109971		S	H			C						H
FUS	89280		S	H	A		c						h

Proteins that copurify with tagged pre-mRNA substrates under conditions that do not support in vitro splicing. Numbered columns same as Table 1.

ding yeast. Perhaps this indicates that mammalian spliceosomes are subject to structural contortions not required in *S. cerevisiae*.

Additional structural rearrangements yet to be characterized might also be suggested by the identification of six new cyclophilin homologs in spliceosomes. Cyclophilins catalyze cis-trans prolyl bond isomerization, facilitating protein conformational changes. Such conformational changes could represent inherently slow steps in the splicing process or function as regulated switches. To date, only one of the spliceosomal cyclophilins, USA-Cyp, has been shown to function in splicing (Horowitz et al., 2002), but its enzymatic target remains unknown. Like the DExD/H proteins, five of the new spliceosomal cyclophilins were found only in mammalian complexes and have no clear homologs in budding yeast.

Auxiliary splicing factors, those proteins that function to recruit the splicing machinery to the fairly degenerate splice sequences were undoubtedly identified in these experiments but were less likely to be found in all complexes due to possibly looser associations and differences in substrates and purification conditions. Included in this class of factors are the SR proteins, which contain a characteristic domain rich in RS dipeptides (Zahler et al., 1992). SR proteins promote both cross-exon and cross-intron interactions between snRNPs,

thereby facilitating spliceosome assembly (Hastings and Krainer, 2001; Zahler et al., 1992). As expected, different complements of SR proteins were found in the different splicing complexes. The few SR proteins indicated in C-complex were identified by only single peptides, and their poor association was attributable to the buffer conditions under which this complex was purified (Reichert et al., 2002).

Proteins that represent links between splicing and other steps in pre-mRNA and mRNA metabolism were also identified in multiple studies. These include components of the polyadenylation machinery and the TREX and exon junction complexes. The polyadenylation proteins found were both subunits of cleavage factor I (CF I) as well as the 160 kDa subunit of the cleavage and polyadenylation specificity factor (CPSF 160). CF I is required for assembly of the 3'-end processing machinery, while CPSF 160 functions in recognizing the polyA site and recruiting polyA polymerase (Shatkin and Manley, 2000). The TREX complex links transcription elongation to mRNA export (Strasser et al., 2002). Its presence in splicing complexes has been suggested to explain how export factors are preferentially loaded onto exons as opposed to introns (Zhou et al., 2002). The exon junction complex is a set of proteins deposited on mRNAs as a consequence of splicing (Le Hir et al., 2000). A number of its components have been shown to function in either mRNA export (Rodrigues et al., 2001; Stutz et al., 2000; Zhou et al., 2000) or nonsense-mediated mRNA decay (Kim et al., 2001; Le Hir et al., 2001; Lykke-Andersen et al., 2000, 2001). Thus, by providing direct evidence of physical associations among the various machineries required to generate cytoplasmic mRNA, the mass spec analyses of spliceosomes further strengthen the case for functional coupling between splicing and other steps in pre-mRNA/mRNA metabolism (Maniatis and Reed, 2002; Proudfoot et al., 2002).

Given recent resurgence of the idea that some translation or reading frame recognition could occur inside the nucleus and possibly even influence pre-mRNA splicing patterns (Wilkinson and Shyu, 2002; although see Dahlberg et al., 2003), one might wonder whether some of the translation machinery detected in spliceosome preps is meaningful rather than merely contaminating. For those ribosomal proteins reported (Makarov et al., 2002; Rappaport et al., 2002; Stevens et al., 2002), the vast majority were from the small subunit. In addition, Makarov et al. (2002) detected several subunits of initiation factor eIF3. In the cytoplasm, eIF3 associates with the small subunit and prevents its reassociation with the large subunit in the absence of a start codon. Do these observations suggest a previously unrecognized link between the spliceosome and the small ribosomal subunit, or do they simply reflect tethering of the small subunit via factors bound to the pre-mRNA 5'-7-methyl-G cap structure? Undoubtedly, future studies will address the functionality of these and other associations.

As for the ~40 remaining proteins with no previous association with pre-mRNA splicing identified multiple times in the mass spec studies, it is hard to say much about them at present. Some of these proteins have domains that one might expect in splicing factors (e.g., homologies to RNA binding proteins or RNA helicases),

but many are classified only as unknown open reading frames with limited homologies to other proteins.

Finally, with all the proteomics data now available for splicing complexes, it is of interest to compare the results with mass spec experiments conducted on an even larger scale. Two papers published last year described systematic affinity-tagging of yeast proteins and identification of copurifying polypeptides by mass spec (Gavin et al., 2002; Ho et al., 2002). The analysis by Gavin et al. included ~30 tagged splicing factors. Although no single splicing factor pulled down a complex that represented the intact spliceosome, the authors presented data analysis suggesting the presence six of distinct complexes primarily populated by spliceosomal proteins. These complexes, designated T1–T6, show patterns of proteins resembling some of the splicing complexes reviewed here (Tables 1 and 2). T2 contains primarily U1 snRNP components, whereas T1 is composed of the Lsms, U5 proteins, and many U4/U6:U5 triple snRNP proteins. T3 tracks most closely to the U2/U5/U6 complex from yeast, while T4, T5, and T6 are all very similar to the penta-snRNP. These last three differ primarily in the nonspliceosomal proteins found in each—for example, a group of mitochondrial ribosomal proteins was proposed to be associated with T6. Interestingly, a subset of the new splicing suspects were also found in these complexes. For example, YLR424W, a *S. cerevisiae* gene of unknown function required for viability and containing a G-patch putative RNA binding motif, is a component of T1 and T3–6. The most highly related human protein, TIP39, was identified in spliceosomes by five different groups. Although TIP39 was originally identified as an interacting partner to a constituent of tooth enamel (Paine et al., 2000), the proteomics data reviewed here strongly suggest its likely function is as a splicing factor. It is nuggets like this that spur prospectors to roll up their sleeves and get back to the lab bench to flesh out the functions of these new splicing factor candidates.

In summary, almost two decades after its discovery, the parts list for the spliceosome is finally nearing completion. However, while obtaining such a comprehensive parts list represents a significant accomplishment in and of itself, what we need now are the detailed assembly instructions. The even greater challenge ahead will be to fit all of these parts into a comprehensive whole in which individual functions and interconnections are defined and understood.

References

- Ajuh, P., Kuster, B., Panov, K., Zomerdiik, J.C., Mann, M., and Lamond, A.I. (2000). Functional analysis of the human CDC5L complex and identification of its components by mass spectrometry. *EMBO J.* 19, 6569–6581.
- Albers, M., Diment, A., Muraru, M., Russell, C.S., and Beggs, J.D. (2003). Identification and characterization of Prp45p and Prp46p, essential pre-mRNA splicing factors. *RNA* 9, 138–150.
- Burge, C.B., Tuschl, T., and Sharp, P.A. (1999). Splicing of precursors to mRNAs by the spliceosomes. In *The RNA World*, Second Edition, R.F. Gesteland, T.R. Cech, and J. F. Atkins, eds. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press), pp. 525–560.
- Chen, C.H., Yu, W.C., Tsao, T.Y., Wang, L.Y., Chen, H.R., Lin, J.Y., Tsai, W.Y., and Cheng, S.C. (2002). Functional and physical interactions between components of the Prp19p-associated complex. *Nucleic Acids Res.* 30, 1029–1037.
- Dahl, R., Wani, B., and Hayman, M.J. (1998). The Ski oncoprotein interacts with Skip, the human homolog of *Drosophila* Bx42. *Oncogene* 16, 1579–1586.
- Dahlberg, J.E., Lund, E., and Goodwin, E.B. (2003). Nuclear translation: what is the evidence? *RNA* 9, 1–8.
- Das, R., Zhou, Z., and Reed, R. (2000). Functional association of U2 snRNP with the ATP-independent spliceosomal complex. *E. Mol. Cell* 5, 779–787.
- Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147.
- Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., and Gygi, S.P. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. USA* 100, 6490–6495. Published online May 27, 2003.
- Gottschalk, A., Neubauer, G., Banroques, J., Mann, M., Luhrmann, R., and Fabrizio, P. (1999). Identification by mass spectrometry and functional analysis of novel proteins of the yeast [U4/U6.U5] tri-snRNP. *EMBO J.* 18, 4535–4548.
- Hartmuth, K., Urlaub, H., Vornlocher, H.P., Will, C.L., Gentzel, M., Wilm, M., and Luhrmann, R. (2002). Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. *Proc. Natl. Acad. Sci. USA* 99, 16719–16724.
- Hastings, M.L., and Krainer, A.R. (2001). Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol.* 13, 302–309.
- Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutillier, K., et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183.
- Horowitz, D.S., Lee, E.J., Mabon, S.A., and Misteli, T. (2002). A cyclophilin functions in pre-mRNA splicing. *EMBO J.* 21, 470–480.
- Huang, T., Vilardell, J., and Query, C.C. (2002). Pre-spliceosome formation in *S.pombe* requires a stable complex of SF1–U2AF(59)–U2AF(23). *EMBO J.* 21, 5516–5526.
- Jurica, M.S., Licklider, L.J., Gygi, S.R., Grigorieff, N., and Moore, M.J. (2002). Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. *RNA* 8, 426–439.
- Kim, V.N., Kataoka, N., and Dreyfuss, G. (2001). Role of the nonsense-mediated decay factor hUpf3 in the splicing-dependent exon-exon junction complex. *Science* 293, 1832–1836.
- Le Hir, H., Moore, M.J., and Maquat, L.E. (2000). Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. *Genes Dev.* 14, 1098–1108.
- Le Hir, H., Gatfield, D., Izaurralde, E., and Moore, M.J. (2001). The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J.* 20, 4987–4997.
- Lykke-Andersen, J., Shu, M.D., and Steitz, J.A. (2000). Human Upf proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. *Cell* 103, 1121–1131.
- Lykke-Andersen, J., Shu, M.D., and Steitz, J.A. (2001). Communication of the position of exon-exon junctions to the mRNA surveillance machinery by the protein RNPS1. *Science* 293, 1836–1839.
- Makarov, E.M., Makarova, O.V., Urlaub, H., Gentzel, M., Will, C.L., Wilm, M., and Luhrmann, R. (2002). Small nuclear ribonucleoprotein remodeling during catalytic activation of the spliceosome. *Science* 298, 2205–2208.
- Maniatis, T., and Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature* 416, 499–506.
- Neubauer, G., Gottschalk, A., Fabrizio, P., Seraphin, B., Luhrmann, R., and Mann, M. (1997). Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl. Acad. Sci. USA* 94, 385–390.
- Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond, A., and Mann, M. (1998). Mass spectrometry

- try and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat. Genet.* 20, 46–50.
- Nilsen, T.W. (2002). The spliceosome: no assembly required? *Mol. Cell* 9, 8–9.
- Ohi, M.D., and Gould, K.L. (2002). Characterization of interactions among the Cef1p-Prp19p-associated splicing complex. *RNA* 8, 798–815.
- Ohi, M.D., Link, A.J., Ren, L., Jennings, J.L., McDonald, W.H., and Gould, K.L. (2002). 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.* 22, 2011–2024.
- Paine, C.T., Paine, M.L., Luo, W., Okamoto, C.T., Lyngstadaas, S.P., and Snead, M.L. (2000). A tuftelin-interacting protein (TIP39) localizes to the apical secretory pole of mouse ameloblasts. *J. Biol. Chem.* 275, 22284–22292.
- Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. *Cell* 108, 501–512.
- Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Seraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* 24, 218–229.
- Ranish, J.A., Yi, E.C., Leslie, D.M., Purvine, S.O., Goodlett, D.R., Eng, J., and Aebersold, R. (2003). The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* 33, 349–355.
- Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Res.* 12, 1231–1245.
- Reichert, V.L., Le Hir, H., Jurica, M.S., and Moore, M.J. (2002). 5' exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. *Genes Dev.* 16, 2778–2791.
- Rodrigues, J.P., Rode, M., Gatfield, D., Blencowe, B.J., Carmo-Fonseca, M., and Izaurralde, E. (2001). REF proteins mediate the export of spliced and unspliced mRNAs from the nucleus. *Proc. Natl. Acad. Sci. USA* 98, 1030–1035.
- Shatkin, A.J., and Manley, J.L. (2000). The ends of the affair: capping and polyadenylation. *Nat. Struct. Biol.* 7, 838–842.
- Staley, J.P., and Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* 92, 315–326.
- Stevens, S.W., and Abelson, J. (1999). Purification of the yeast U4/U6.U5 small nuclear ribonucleoprotein particle and identification of its proteins. *Proc. Natl. Acad. Sci. USA* 96, 7226–7231.
- Stevens, S.W., Ryan, D.E., Ge, H.Y., Moore, R.E., Young, M.K., Lee, T.D., and Abelson, J. (2002). Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol. Cell* 9, 31–44.
- Strasser, K., Masuda, S., Mason, P., Pfannstiel, J., Oppizzi, M., Rodriguez-Navarro, S., Rondon, A.G., Aguilera, A., Struhl, K., Reed, R., and Hurt, E. (2002). TREX is a conserved complex coupling transcription with messenger RNA export. *Nature* 417, 304–308.
- Stutz, F., Bachi, A., Doerks, T., Braun, I.C., Seraphin, B., Wilm, M., Bork, P., and Izaurralde, E. (2000). REF, an evolutionary conserved family of hnRNP-like proteins, interacts with TAP/Mex67p and participates in mRNA nuclear export. *RNA* 6, 638–650.
- Tam, W.Y., Hsu, C.H., Huang, K.T., Chen, H.R., Kao, H.Y., Lee, K.R., and Cheng, S.C. (1994). Functional association of essential splicing factor(s) with PRP19 in a protein complex. *EMBO J.* 13, 2421–2431.
- Verma, R., Chen, S., Feldman, R., Schieltz, D., Yates, J., Dohmen, J., and Deshaies, R.J. (2000). Proteasomal proteomics: identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. *Mol. Biol. Cell* 11, 3425–3439.
- Wang, Y., and Rando, R.R. (1995). Specific binding of aminoglycoside antibiotics to RNA. *Chem. Biol.* 2, 281–290.
- Wilkinson, M.F., and Shyu, A.B. (2002). RNA surveillance by nuclear scanning? *Nat. Cell Biol.* 4, E144–147.
- Will, C.L., Urlaub, H., Achsel, T., Gentzel, M., Wilm, M., and Luhrmann, R. (2002). Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. *EMBO J.* 21, 4978–4988.
- Xu, D., and Friesen, J.D. (2001). Splicing factor slt11p and its involvement in formation of U2/U6 helix II in activation of the yeast spliceosome. *Mol. Cell. Biol.* 21, 1011–1023.
- Zahler, A.M., Lane, W.S., Stolk, J.A., and Roth, M.B. (1992). SR proteins: a conserved family of pre-mRNA splicing factors. *Genes Dev.* 6, 837–847.
- Zhou, Z., Licklider, L.J., Gygi, S.P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. *Nature* 419, 182–185.
- Zhou, Z., Luo, M.J., Straesser, K., Katahira, J., Hurt, E., and Reed, R. (2000). The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature* 407, 401–405.