Analysis

Pre-messenger RNA Processing Factors in the *Drosophila* Genome

Stephen M. Mount* and Helen K. Salz‡

*Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, College Park, Maryland 20742-5815; and ‡Department of Genetics, Case Western Reserve University, Cleveland, OH 44106-4955

In eukaryotes, messenger RNAs are generated by a process that includes coordinated splicing and 3′ end formation. Factors essential for the splicing of mRNA precursors (pre-mRNA)[1] in eukaryotes have been identified primarily through the study of nuclear extracts derived from mammalian cells and *Saccharomyces cerevisiae* genetics. Here, we identify homologues of most known pre-mRNA processing factors in the recently completed sequence of the *Drosophila* genome. The set of proteins required for RNA processing shows remarkably little variation among eukaryotic species, and individual proteins are highly conserved. In general, proteins involved in the mechanics of RNA processing are even more conserved than proteins involved in the interpretation of RNA processing signals. The genome does not appear to contain a gene for the U11 RNA, or for a protein unique to the U11 snRNP, which raises the possibility that the U12-dependent spliceosome functions without U11 in *Drosophila*.

Introduction

Most RNA processing factors have been identified in either nuclear splicing extracts derived from mammalian cells or in *Saccharomyces cerevisiae* (Burge et al., 1999; Kambach et al., 1999; Minvielle-Sebastia and Keller, 1999). However, *Drosophila* is extensively used for genetic investigations of complex and regulated splicing. In this review, we survey the recently complete *Drosophila* sequence (Adams et al., 2000) for sequences related to factors identified in these other systems. In many cases, functional data for the *Drosophila* protein are not available, and our assignments are based on the best match among genomes. We have not included genes that have been identified in *Drosophila* for which there is evidence of a role in splicing (see, for example, the list presented in Burnette et al., 1999). This analysis yields a list of 27 genes that encode small nuclear RNAs (snRNAs; see Table I) and a list of 99 genes that encode proteins involved in RNA processing (see Table II). Our survey confirms that the components of the RNA processing machinery are highly conserved. Very few factors identified in other species are absent from the *Drosophila* genome. In general, the *Drosophila* proteins are more closely related to their vertebrate counterparts than to the *Saccharomyces cerevisiae* proteins.

Address correspondence to Stephen M. Mount, Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, College Park, MD 20742-5815. Tel.: (301) 405-6934. Fax: 301-314-9081. E-mail: sm193@umail.umd.edu

[1]*Abbreviations used in this paper:* hnRNA, heterogeneous RNA; pre-mRNA, precursors to mRNA; snRNA, small nuclear RNA; snRNP, small nuclear ribonucleoprotein.

Methods

Protein sequences of known yeast and human splicing factors were used to query the annotated set of predicted *Drosophila* proteins using BLASTP, and the nucleotide sequence of the genome using tblastn, on the NCBI server (Altschul et al., 1997; http://www.ncbi.nlm.nih.gov/). All identified *Drosophila* genes were used to query the nonredundant database to establish the optimal yeast and human matches. Alignments were generated using the blast two sequences option (Tatusova et al., 1999) or LALIGN (Huang and Miller, 1991). Cytological positions were taken from GadFly (http://hedgehog.lbl.gov:8000) or flybase (http://flybase.bio.indiana.edu/), or deduced from the positions of flanking genes.

To identify snRNA genes, the *Drosophila* genome was queried using modified blastn parameters (parameter set A: -r 10 -q -11 -W 8 -G 100 -E 50; B: -r 10 -q -11 -W 7 -G 5 -E 20; C: -r 10 -q -11 -W 7 -G 15 -E 4; D: -r 7 -q -14 -W 7 -G 7 -E 3; and E: -r 4 -q -5 -W 8 -G 10 -E 2).

A curated database containing these results will be available at http://www.wam.umd.edu/∼smount/DmRNA factors/table.html.

Results and Discussion

Major and minor snRNP components

U snRNAs. Two types of spliceosomes have been previously described (Burge et al., 1999). The more common U2-type spliceosome is responsible for splicing the majority of introns, and the U12-type spliceosome is responsible for splicing a minor class of rare introns (perhaps 0.1% in both humans and flies). The *Drosophila* genome contains multiple copies of the 5 U snRNAs found in the major class spliceosomes. We found five genes for U1, six genes for U2, three genes for U4, seven genes for U5, and three genes for U6 (Table I). With the exception of U4-25F, and the U5 genes (which were previously known only by in situ hybridization), these genes had been described previously (Alonso et al., 1984; Das et al., 1987; Saba et al., 1986; Saluz et al., 1988; Lo and Mount, 1990). The variant U4-25F has only 69% identity with the major form of fly U4 (Saba

Table I. Genes for Drosophila snRNAs Involved in Splicing

Drosophila gene	Accession	Cytological position
U1a-21D	AE003588	21D
U1a-82Eb	X53542	82E
U1a-82Ec	AE003604	82E
U1a-95Ca	AE003745	95C
U1a-95Cb	AE003745	95C
U1c-95Cc	AE003745	95C
U2-14B	AE003501	14B
U2-34ABa	AE003639	34A-34B
U2-34ABb	AE003639	34A-34B
U2-34ABc	AE003639	34A-34B
U2-38ABa	AE003664	38A5-38B4
U2-38ABb	AE003664	38A5-38B4
U4-38AB	AE003664	38A5-38B4
U4-39B	AE003669	39B1-39B3
U4-25F‡	AE003610	25F
U5-14B	AE003501	14B
U5-23D	AE003581	23D
U5-34A	AE003639	34A-34B
U5-35D	AE003648	35D4
U5-38ABa	AE003664	38A5-38B4
U5-38ABb	AE003664	38A5-38B4
U5-63BC	AE003477	63BC
U6-96Aa	AE003748	96A
U6-96Ab	AE003748	96A
U6-96Ac	AE003748	96A
U4atac-82E‡	AE003603	82E
U6atac-29B‡	AE003621	29B
U12*	AE003526	73B

*Although the human U12 is only 150 nucleotides, the fly U12 RNA is 238 nucleotides and the alignment, although fragmentary, includes both ends of the RNA.
‡Hypothetical. No published data on *Drosophila* RNA supports these assignments.

et al., 1986), and 68% with human U4. Although the possibility that these new snRNA genes are pseudogenes cannot be ruled out, they appear likely to be functional because of their highly conserved promoters. In the case of U4-25F, some of the variation includes compensatory changes that allow formation of conserved stem loop structures. There are four clusters of snRNA genes, including one at 38AB with two U2, one U4, and two U5 genes within 6 kb.

The *Drosophila* genome also contains introns that resemble the minor class (or U12) introns first identified in mammals (Adams et al., 2000). These are recognized by the U12-type spliceosome including U11, U12, U4atac, and U6atac snRNAs in place of U1, U2, U4, and U6 (Hall and Padgett, 1994; Tarn and Steitz, 1996). Identification of snRNAs for the U12-type spliceosome in the genome involved modification of the standard parameters for BLASTN (see Methods). It was possible to find one gene for U12 snRNA, one gene for U6atac snRNA, and one gene for U4atac snRNA. These are almost certainly authentic genes, as critical sequences are conserved. In addition, the highly conserved snRNA promoter is present in each case, including a 9/10 or perfect match to the PSE consensus TAATTCCCAA, which is ~52 nucleotides upstream of the start (Jensen et al., 1998; Lo and Mount, 1990). In contrast, no gene for the U11 snRNA was found. Consistent with the absence of a U11 snRNA, we also failed to find the U11 35-kD–specific protein (accession No. NP_008951; Will et al., 1999). In fact, the U11 snRNP, which functions in 5′ splice site recognition, may not be required for splicing. The highly conserved minor class 5′ splice sites could be recognized by an unknown protein that acts during the early steps of splicing, by the U6atac snRNA alone, or by both. This mechanism would be analogous to a situation seen in vitro, where certain vertebrate introns can be processed in the absence of U1 snRNP if the 5′ splice sites can be recognized by U6 snRNA (Crispino et al., 1994; Tarn and Steitz, 1994).

snRNP Proteins. Each snRNP contains a set of Sm core proteins shared with the other snRNPs and a set of proteins that are specific to that snRNP. All 15 known proteins of the Sm family were identified and are highly conserved. These include seven Sm proteins that bind to the U1, U2, U4, and U5 snRNPs (B, D1, D2, D3, E, F, and G); the seven related LSm proteins found in the U6 snRNP (LSM2-LSM8); and the CaSm/LSM1p protein (Bouveret et al., 2000; Tharun et al., 2000). These and subsequent matches are shown in Table II. The table reports, for each *Drosophila* gene, the GenBank accession number and expectation value (the expected number of matches this good or better; Altschul et al., 1997) for the best human and yeast (*Saccharomyces cerevisiae*) match.

In addition to the Sm proteins, each snRNP also contains a set of snRNP-specific proteins. As expected, orthologues of the proteins that are contained in both the vertebrate and *Saccharomyces cerevisiae* snRNPs are easily identified in *Drosophila,* based on their extensive sequence homology, except that a single *Drosophila* protein, encoded by the *sans-fille (snf)* gene, corresponds to both the U1 snRNP-U1A and U2 snRNP-U2B′′ proteins (Polycarpou-Schwarz et al., 1996; Stitzinger et al., 1999), and no additional homologues were found.

Interestingly, the *Saccharomyces cerevisiae* U1 snRNP is more complex than the vertebrate U1 snRNP, with seven additional protein components that are not found in the purified vertebrate U1 snRNP (Gottschalk et al., 1998; Rigaut et al., 1999). Only two of these proteins, Luc7 and Prp40, have easily identifiable *Drosophila* orthologues. The *Drosophila* ortholog of Luc7, CG7564, is 33% identical to the entire Luc7 protein (Fortes et al., 1999), and a second *Drosophila* Luc7-related protein is 21% identical to the yeast protein (Fortes et al., 1999). We identified a single Prp40-like gene in the *Drosophila* database. CG3542 shares 23% identity with the entire yeast Prp40 protein and 41% sequence identity over its entire length of 757 amino acids with the human protein, FBP11. FBP11 was initially identified because it also contains a tyrosine-rich WW domain and like Prp40, it interacts with the splicing factor SF1 (Bedford et al., 1997). These observations suggest that the function of these proteins in forming bridges between 5′ splice sites and the branchpoint may be conserved (Abovich and Rosbash, 1997). It is likely that these *Drosophila* proteins, like their human homologues, would not be found in purified U1 snRNPs, but, nevertheless, do share a function with their yeast counterparts. A second human Prp40-like protein, FBP21, has been described in the literature (Bedford et al., 1998). FBP21 is more closely related to the *Drosophila* CG4291, with 28% identity over the entire length of the 338–amino acid protein. Because similarity between FBP11/CG3542 and FBP21/CG4291 is limited to the WW repeats, FBP21/

Table II. Genes for Drosophila Proteins Involved in RNA Processing

Drosophila gene	Accession No.	Protein	Cytology	Human homologues, E value	Yeast (S.c.), E value
snRNP proteins					
Core proteins					
CG5352	AAF52947	SmB	31E	NP_003082, 2E-38	SMB1, NP_010946, 2E-10
CG10753	AAF49893	SmD1	69D	NP_008869, E-36	SMD1, NP_011588, 5E-11
CG1249	AAF54135	SmD2	83E5	NP_004588, 9E-46	SMD2, NP_013377, 5E-25
SmD3 (CG8427)	AAF58566	SmD3	48F3	NP_004166, E-37	SMD3, NP_013248, E-20
CG18591	AAF52576	SmE	28D7	NP_003085, 8E-33	SME1, NP_014802, E-17
DebB (CG16792)	AAF58559	SmF	48F3	NP_003086, 6E-37	SMX3, NP_015508, 3E-16
CG9742	AAF48634	SmG	14F6	NP_003087, 7E-27	SMX2, NP_011170, 7E-16
CG4279	AAF46688	LSM1	57C3	CAB45865, E-43	LSM1, NP_01241, 7E-19
CG10418	AAF49929	LSM2	69BC	AAD56226, 3E-46	LSM2, NP_009527, 8E-28
CG5926	AAF56219	LSM3	95D7	CAB45866, 6E-29	LSM3, NP_013543, E-05
CG17764	AAF52939	LSM4	31B	NP_036453, 3E-70	LSM4, NP_011037, E-10
CG6610	AAF50703	LSM5	65A6	NP_036454, 5E-35	LSM5, NP_011073, E-13
CG9344	AAF46645	LSM6	57B11	NP_009011, 8E-33	LSM6, NP_010666, 7E-09
CG13277	AAF53562	LSM7	36A8	AAD56231, 8E-38	LSM7, NP_014252, 3E-16
CG2021	AAF47567	LSM8	62AB	AAD56232, E-34	LSM8, NP_012556, E-05
U1 snRNP					
snRNP70K (CG8749)	AAF52471	U1-70k	27C9	NP_003080, 2E-62	Snp1, NP_012203, 7E-31
CG5454	AAF55616	U1C	91E2	NP_003084, 8E-28	Yhc1, NP_013401, 4E-06
snf (CG4528)*	P43332	U1A	4F2	NP_004587, 7E-32	Mud1, NP_009677, 6E-04
CG7564‡	AAF49330	Luc7-like	74C1	BAA91737, 4E-84	Luc7, NP_010196, 5E-24
CG3198‡	AAF46183	Related to Luc7,	6C8	BAA90542, 3E-06	Luc7, NP_010196, 3E-06
CG1646‡	AAF56816	Prp39/Prp42-like	98F4	BAA92024, 2E-30; BAA91318, 9E-23	Prp39, NP_013667, 2E-13; Prp42, NP_010521, 1E-06
CG3542	AAF51161	Prp40-like	23C	FBP11, AAD39463, 2E-98	Prp40, NP_012913, 3E-14
Rox8, CG5422	AAF56225	Similar to Nam8 See text	95D	NP_003243, E-82	Pub1p, NP_014382, E-28 Nam8, NP_011954, E-19
U2 snRNP					
snf (CG4528)*	P43332	U2B′′	4F2	NP_003083, 7E-68	Yib9, NP_012274, 3E-05
CG1406	AAF59207	U2A′	43E8	NP_003081, 7E-54	Lea1, NP_015111, 5E-05
noi (CG2925)	CAA11045	SF3a60/SAP61	83B4	NP_006793, E-180	Prp9, NP_010254, 3E-31
CG10754	AAF49876	SF3a66/SAP62	69E2	NP_038679, E-100	Prp11, NP_010241, 1E-13
CG16941	AAF55372	SF3a120/SAP114	89E3-4	NP_005868, E-155	Prp21, NP_012332, 2E-11
Spx (CG3780)	AAF46136	SF3b53/SAP49	5E1	NP_005841, E-109	Hsh49, NP_014964, 1E-28
CG3605	AAF51159	SF3b150/SAP145	23C	NP_006833, E-178	Cus1, NP_013967, 5E-38
CG13900	AAF47416	SF3b120/SAP130	61D4-E1	CAB56791, E = “0”	Rse1, NP_013663, 8E-78
CG2807	AAF51478	SF3b160/SAP155	21C6	NP_036565, E = “0”	Ymr288w, NP_014015, E = “0”
CG4291	AAF51446	FBP21-like	21D2	FBP21, AAC34810, 7E-28	Prp40, NP_012913, 0.078
U5 snRNP					
CG8877	AAF58573	220 kD	48F1	NP_006436, E = “0”	Prp8, NP_012035, E = “0”
CG5205	AAF55204	200 kD	88F8	CAA94089, E = “0”	Ygr271w, NP_011787, E = “0”; Brr2, NP_011099, E = “0”
CG3436	AAF51536	40 kD	21B2	NP_004805, E-120	unknown
CG4849	AAF56769	116 kD	98B8	NP_004238, E = “0”	Snu114, NP_012748, E-156
CG6841	AAF49211	Prp6-like	75E1	BAA37140, E = “0”	Prp6, NP_009611, 6E-74
CG10333	AAF53680	100 kD	37A4	NP_004809, E = “0”	Prp28, NP_010529, 2E-92
CG3058	AAF51017	15 kD	25A1	NP_006692, 5E-78	Dib1, NP_015407, 3E-48
TrisnRNP					
CG7757	AAF49097	SAP90	80h	NP_004689, E-144	Prp3, NP_010761, 1E-23
CG6322	AAF49331	SAP60	74C1	NP_004688, E-138	Prp4, NP_015504, 9E-56
Hoi-Polloi (CG3949)	AAF20209	15.5 kD	30C7	NP_004999, 4E-39	Snu13, NP_010888, 9E-32
CG17266	AAF57375	20 kD cyclophilin	42C5	NP_006338, 4E-80	unknown

(Table II continues)

Table II. (continued)

Drosophila gene	Accession No.	Protein	Cytology	Human homologues, E value	Yeast (S.c.), E value
Prespliceosome proteins					
U2af38 (CG3582)	Q94535	U2AF small subunit	21B7	NP_006749, 5E-88	unknown
CG3294	AAF50982	Related to U2AF small subunit	25B4-B5	BAA08533, 1E-20 U2AF1-RS2	unknown
U2af50 (CG9998)	Q24562	U2AF large subunit	14C1	NP_009210, E-146	Mud2, NP_012849, 0.22
CG5836	CAB64937	SF1/BBP	90B1	CAA03883, E-106	Bbp1, Msl5, NP_013217, 4E-62
CG12357	AAF55496	CBP20	90E3-4	CBP20, NP_031388, 4E-63	Mud13, NP_015147, 6E-37
Cbp80 (CG7035)	AAF45970	CBP80	4C9-10	CBP80, NP_002477, E= “0”	GCR3, NP_013844, 3E-23
CG7907	AAF55893	CBP80-like	36C	CBP80, NP_002477, 1E-96	GCR3, NP_013844, 1E-10
Hel25E (CG7269)	AAF52261	UAP56	25F1	NP_004631, E = “0”	Sub2, NP_010199, E-141
CG6227	AAF48446	Prp5-like	13C5	NP_055644, E = “0”	Prp5, NP_009796, 4E-94
Catalytically active spliceosome proteins					
CG6876	AAF49655	Prp31-like	71B2	CAB43677, E-141	Prp31, NP_011605, 3E-22
CG6905	AAF47383	pombe, cdc5-related	61C	AAB61210, E = “0”	Cef1, NP_013940, 4E-49
CG10689	AAF53766	Prp2-like	37C	NP_003578, E = “0”	Prp2, NP_014408, E-172
CG8241	AAF58294	Prp22-like	50E	HRH1, NP_004932, E = “0”	Prp22, NP_010929, E = “0”
CG1405	AAF48355	Prp16-like	12D1	NP_054722, E = “0”	Prp16, NP_013012, E-160
CG6015	AAF55949	Prp17-like	93F14	AAC39730, E = “0”	Prp17, NP_010652, 1E-76
CG1420	AAF56845	Slu7-like	98F13	NP_006416, E-128	Slu7, NP_010373, 1E-12
Prp18 (CG6011)	AAF55627	Prp18-like	91E4	NP_003666, 4E-93	Prp18, NP_011520, 4E-05
SR proteins					
SC35 (CG5442)	AAF53192	SC35	33DE	SC35, NP_035488, 5E-53	Npl3, NP_010720, 3E-12
ASF/SF2 (CG6987)	AAF55300	ASF/SF2	89C	ASF/SF2, NP_008855, 2E-80	Npl3, NP_010720, 7E-15
B52 (CG10851)	AAF54968	B52	87F	SRp55, CAB43960, 2E-96; SRp75, NP_005617, 3E-96; Srp40/HRS, Q13243, E-78	Npl3, NP_010720, 3E-14
9G8 (CG10203)	AAF52454	9G8	27C	9G8, NP_006267, 4E-42	Npl3, NP_010720, 7E-05
RBP1 (CG17136)	AAF54555	RBP1	86C12	SRP20, NP_003008, 2E-38	Nsr1, NP_011675, E-07
RBP1-like (CG1987)	AAF48264	RBP1-like	11F3	SRP20, NP_003008, 4E-31	Nsr1, NP_011675, 4E-14
SRp54 (CG4602)	AAF52825	SRp54	30DE	SRp54, NP_004759, 2E-41	unknown
SR protein kinases					
CG8174	AAF58140	SRPKD1	52A1	SRPK1, NP_003128, 9E-76; SRPK2, NP_003129, 3E-77	Sky1, NP_0013943, 5E-44
CG9085	AAF51819	SRPKD2	79E4	SRPK1, NP_003128, 2E-63 SRPK2, NP_003129, 3E-66	Sky1, NP_0013943, 3E-49
CG8565	AAF48523	SRPKD3	13F3-4	SRPK1, NP_003128, 8E-50; SRPK2, NP_003129, 3E-50	Sky1, NP_0013943, 2E-70
Doa (CG1658)	AAF56832	LAMMER/CLK kinase	98F6	CLK2, NP_003984, E-151	Kns1, NP_013081, 5E-61
Miscellaneous proteins					
Crn (CG3193)	AAF45760	crn/CLF1	2F1	NP_057736, E = “0”	Clf1, NP_013218, E-123
Puf60 (CG12085)	AAF47501	PUF60 (poly-U binding)	62A	AAF05605, 1E-81	unknown
nonA-1 (CG10328)	AAF55347	PSF	89C7	NP_005057, 5E-62	unknown
PTB (CG2094)	AAF57208	PTB/hnRNP I	100F	NP_002810, E-150	unknown
CG11107	AAF59269	Prp43-like	43B2	NP_001349, E = “0”	Prp43, NP_011395, E = “0”
CG11274	AAF49848	Srm160-like	69F2	NP_005830, 2E-43	unknown

(Table II continues)

Table II. (continued)

Drosophila gene	Accession No.	Protein	Cytology	Human homologues, E value	Yeast (S.c.), E value
CG7971	AAF47543	Srm300-like	62B4	AAF21439, 2E-37	unknown
CG16725	AAF49446	Similar to SMN	73A5	NP_035550, 1E-05	unknown
CG17454	AAF45352	SPF30 SMN-related	80C-82B	NP_005862, 3E-37	unknown
CG10419	AAF49187	SMN-interacting (SIP1)	75F3	NP_003607, 2E-18	Brr1, NP_015382, E = 0.05
CG7942	AAF50486	Debranching enzyme	66B9-10	AAD53327, E-113	Dbr1/Prp26, NP_012773, 2E-62
CG11266	AAF52478	CC1.3; splicing factor domains only	27D3	NP_004893, E-123	unknown
BcDNA:GH01073 (UKL; CG8108)	AAF56342	Splicing factor domains only	96B1	AC023603, E-39	unknown
RSF1 (CG5655)	AAF52890	Repressive splicing factor	31C5	unknown	unknown
Cleavage and polyadenylation					
General					
CG9854	AAF57528	PolyA polymerase	56D11	P51003, E-173	Pap1, P29468, E-102
PAbp (CG5119)	AAF57747	PolyA binding protein	55B9-10	P29341, E-176 (mouse)	Pab1, AAA34838, E-106
Pabp2 (CG2163)	AAF59127	PolyA binding protein II	44B	NP_004634, 2E-54	Sgn1, NP_012266, 2E-18
CG4612	AAF47219	Possible polyA binding protein	60D	CAA15498, 8E-46	Pab1, AAA34838, 2E-24
CPSF					
CG10110	AAF58240	CPSF-160 kD	51A7	Q10570, E = "0"	Cft1, NP_010587, 4E-37
BcDNA:LD14168 (CG1957)	AAF56844	CPSF-100 kD	98F13	Q10568, E = "0", bovine	Ysh1, NP_013379, E-32
CG7698	AAF55578	CPSF-73 kD	91B7	AAB70268, E = "0"	Ysh1, NP_013379, E-146
CG1972	AAF56931	CPSF-73–kD variant	99C2	AAF00224, 7E-93	Ysh1, NP_013379, E-73
CG5222	AAF49538	Related to CPSF-100 and -73	72D5	BAA91867, E = "0"; AAB67601, E-113	Ysh1, NP_013379, 6E-14
Clp (CG3642)	AAF51453	CPSF-30 kD	21D2	AAD00321, 8E-89	Yth1, NP_015432, 1E-34
CstF					
Su(f) (CG17170)	AAF45314	CstF-77 kD	20E	NP_001317, E = "0"	Rna14, NP_013777, 2E-47
CstF-64 (CG7697)	AAF55577	CstF-64 kD	91B7	NP_001316, E-71	Rna15, NP_011471, 2E-18
CstF-50 (CG2261)	AAF57183	CstF-50 kD	100E1-2	NP_001315, E-129	unknown
CG2097	AAF51962	Symplekin	83C1	NP_004810, E-173	Pta1, NP_009356, 5E-04
Cleavage factors I & II					
CG3689	AAF50278	25-kD subunit	67A6	NP_008937, E-102	unknown
CG7185	AAF50445	68-kD subunit	66C5	NP_008938, 3E-34	unknown
CG10228	AAF46699	Factor IA	51D2-4	KIAA0824, 7E-41	Pcf11, NP_010514, 3E-9

**snf* is listed twice in the table because it is the counterpart of both the U1 snRNP-U1A and U2 snRNP-U2B′′ proteins.
‡Categorized as a U1snRNP protein based on homology with the yeast U1snRNP proteins (see text for details).

CG4291 is unlikely to be related to Prp40. Consistent with this idea, human FBP21 has been found to stably associate with the U2 snRNPs and, therefore, may function at a later stage of spliceosome assembly than does Prp40 (Bedford et al., 1998).

Searches with the yeast U1 snRNP proteins Prp39 and Prp42 have identified only a single homologous sequence. Prp39 and Prp42 belong to a family of TPR repeat proteins (McLean and Rymond, 1998) and share 25% sequence identity with each other over a ~270–amino acid region that includes several copies of the TPR repeat motif. We identified a single *Drosophila* protein, encoded by the CG1646 gene, that shares 25% sequence identity with Prp39 and Prp42 over the same ~270 amino acids, and is the best match between the *Saccharomyces cerevisiae* and *Drosophila* genomes. The *Drosophila* crooked neck protein (Crn) is another TPR repeat protein, it's yeast homologue has been shown to act later in spliceosome assembly (Chung et al. 1999).

Surprisingly, there are three *Saccharomyces cerevisiae* U1 snRNP proteins that have no clear counterparts in the *Drosophila* database. No *Drosophila* proteins, whose best match in the *S. cerevisiae* genome is Snu71, Snu56 or Nam8, were found. Recent work on Snu56 and Nam8 suggests that these proteins contact the pre-mRNA directly and may anchor the U1 snRNP onto the substrate (Puig et al., 1999; Zhang and Rosbash, 1999), a function that could be dispensable in metazoans because it could be provided by the SR proteins. Alternatively, a similar function may be provided by proteins, such as *Drosophila* rox8 in the case Nam8, that do not appear to be orthologues (*Drosophila* rox8 matches three other yeast proteins better than Nam8). Proteins in the U2 snRNP, U5 snRNP and U4/U6.U5 tri-snRNP are generally very conserved, and no

significant differences between *Drosophila* and other species were revealed by our analysis (Table II).

Proteins Required for Splice Site Selection

SR proteins are splicing factors that contain either one or two characteristic RNA-binding domains and an RS domain. These proteins are among the earliest acting proteins in spliceosome assembly (Zahler et al., 1992; Graveley et al., 1999; Tacke and Manley, 1999). There are 11 well characterized mammalian SR proteins: 9G8, SRp20, ASF/SF2, SC35, SRp30c, p54, SRp40, SRp55, SRp75 (for review see Mount, 1997), NSSR1, and NSSR2 (Komatsu et al., 1999). Individual SR proteins differ with respect to the sequence specificity of their RNA-binding domains, and with respect to their ability to recognize and activate different exonic splicing enhancer sequences. We have identified seven SR protein genes in the *Drosophila* genome. These include the previously described B52, RBP1, SRp54 (Kennedy et al., 1998) and X16/9G8 (Vorbruggen et al., 2000) genes, as well as *Drosophila* orthologues of ASF/SF2 and SC35. In addition, we have identified a novel gene, CG1987, that is 95% identical to RBP1.

Phosphorylation of SR proteins is thought to play an important role in controlling spliceosome assembly (Stojdl and Bell, 1999; Yeakley et al., 1999). Both SRPK and LAMMER (or CLK) kinases phosphorylate SR proteins. We have identified three kinases of the SRPK type (CG8174, CG9085, and CG8565) and only one LAMMER kinase, the previously described Doa kinase (Du et al., 1998).

A variety of proteins bind to pre-mRNA (also known as hnRNA), and many of these proteins, defined as hnRNP proteins, have been shown to influence splicing, typically by inhibition of splicing events near their binding sites (Chen et al., 1999). A number of *Drosophila* hnRNP proteins have been described (e.g., Matunis et al., 1992), and some nuclear RNA-binding proteins without clear homologues in mammalian species have unambiguous roles in the regulation of splicing (e.g., SXL). However, because it is impossible to determine from sequence alone whether a given RNA-binding protein is likely to function in splicing, or even to reside in the nucleus, we have not undertaken an analysis of these proteins. These proteins are discussed in the accompanying article by Lasko (2000).

Genome Contents: Parallels and Differences

The results of our search for RNA processing factors known from studies in mammalian extracts and *Saccharomyces cerevisiae* genetics indicate that very few RNA processing factors are absent from the *Drosophila* genome. Indeed, our survey reveals remarkably little variation in this list among yeast, flies, and mammals. As expected, *Drosophila* proteins are more closely related to their vertebrate counterparts than to the *Saccharomyces cerevisiae* proteins.

The extensive conservation of the components of the spliceosome between vertebrates and *Drosophila* supports the suggestion that the primary mode of regulating splicing takes place at the level of spliceosome assembly (Lopez, 1998; Staley and Guthrie, 1998). Some of these factors, such as SR proteins, which regulate assembly of the spliceosome on many different RNAs, are well conserved and are easily identifiable. Missing from these tables are the factors that regulate the splicing of specific RNAs. These factors are less likely to be well conserved and, indeed, some may prove to be organism specific (e.g., SXL and TRA). Even the variation we observe among proteins and RNAs with clearly established roles in splicing, per se, is weighted towards the early events in splice site selection. For example, the *Drosophila* genome is missing a set of U1 snRNP proteins that are found in the yeast U1 snRNP but not in the vertebrate U1 snRNP, and the genome does not appear to contain a gene for the U11 RNA, or for the single known protein unique to the U11 snRNP, suggesting that U12 functions without U11 in *Drosophila.* Here again, variation is observed in components that function in splice site selection.

We thank Jonathan Roberts (University of Maryland, College Park, MD) for help automating the web searches and for help with the website. We thank Jo Ann Wise (Case Western Reserve University, Cleveland, OH), Javier Lopez (Carnegie Mellon University, Pittsburgh, PA), and Jim Manley (Columbia University, New York, NY) for discussions and comments on the manuscript. We apologize to those whose relevant publications could not be cited due to space limitations.

This work was supported by GM37991-11 to SMM and NSF-MCB9904565 to H.K. Salz.

Submitted: 30 May 2000
Revised: 22 June 2000
Accepted: 22 June 2000

References

Abovich, N., and M. Rosbash. 1997. Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell.* 89:403–412.

Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, and J.C. Venter. 2000. The genomic sequence of *Drosophila melanogaster. Science.* 287:2185–2195.

Alonso, A., E. Beck, J.L. Jorcano, and B. Hovemann. 1984. Divergence of U2 snRNA sequences in the genome of D. melanogaster. *Nucleic Acids Res.* 12: 9543–9550.

Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Bedford, M.T., D.C. Chan, and P. Leder. 1997. FBP WW domains and the Abl SH3 domain bind to a specific class of proline-rich ligands. *EMBO (Eur. Mol. Biol. Organ.) J.* 16:2376–2383.

Bedford, M.T., R. Reed, and P. Leder. 1998. WW domain-mediated interactions reveal a spliceosome-associated protein that binds a third class of proline-rich motif: The proline glycine and methionine-rich motif. *Proc. Natl. Acad. Sci. USA* 95:10602–10607.

Bouveret, E., G. Rigaut, A. Shevchenko, M. Wilm, and B. Seraphin. 2000. A Sm-like protein complex that participates in mRNA degradation. *EMBO (Eur. Mol. Biol. Organ.) J.* 19:1661–1671.

Burge, C.B., T. Tuschl, and P.A. Sharp. 1999. Splicing of precursors to mRNAs by the spliceosome. *In* The RNA World. R.F. Gesteland and J.F. Atkins, editors. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. 525–560.

Burnette J., A.R. Hatton, and A.J. Lopez. 1999. Trans-acting factors required for inclusion of regulated exons in the ultrabithorax mRNAs of *Drosophila melanogaster. Genetics* 151:1517–1529.

Chen, C.D., R. Kobayashi, and D.M. Helfman. 1999. Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. *Genes Dev.* 13:593–606.

Chung, S., M.R. McLean, and B.C. Rymond. 1999. Yeast ortholog of the *Drosophila* crooked neck protein promotes spliceosome assembly through stable U4/U6.U5 snRNP addition. *RNA* 5:1042–1054.

Crispino, J., B.J. Blencowe, and P.A. Sharp. 1994. Complementation by SR proteins of pre-mRNA splicing reactions depleted of U1 snRNP. *Science.* 2:664–673.

Das, B., D. Henning, and R. Reddy. 1987. Structure, organization, and transcription of *Drosophila* U6 small nuclear RNA genes. *J. Biol. Chem.* 262: 1187–1193.

Du, C., M.E. McGuffin, B. Dauwalder, L. Rabinow, and W. Mattox. 1998. Protein phosphorylation plays an essential role in the regulation of alternative splicing and sex determination in *Drosophila. Mol. Cell.* 2:741–750.

Fortes, P., D. Bilbao-Cortes, M. Fornerod, G. Rigaut, W. Raymond, B.

Seraphin, and I.W. Mattaj. 1999. Luc7p, a novel yeast U1 snRNP protein with a role in 5′ splice site recognition. *Genes Dev.* 13:2425–2438.

Gottschalk, A., J. Tang, O. Puig, J. Salgado, G. Neubauer, H.V. Colot, M. Mann, B. Seraphin, M. Rosbash, R. Luhrmann, and P. Fabrizio. 1998. A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins. *RNA*. 4:374–393.

Graveley, B.R., K.J. Hertel, and T. Maniatis. 1999. SR proteins are locators of the RNA splicing machinery. *Curr. Biol*. 9:R6–R7.

Hall, S.L., and R.A. Padgett. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J. Mol. Biol.* 239: 357–365.

Huang, X., and W. Miller. 1991. A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.* 12:373–381.

Jensen, R.C., Y. Wang, S.B. Hardin, and W.E. Stumph. 1998. The proximal sequence element (PSE) plays a major role in establishing the RNA polymerase specificity of *Drosophila* U-snRNA genes. *Nucleic Acids Res.* 26: 616–622.

Kambach, C., S. Walke, and K. Nagai. 1999. Structure and assembly of the spliceosomal small nuclear ribonucleoprotein particles. *Curr. Opin. Struct. Biol.* 9:222–230.

Kennedy, C.F., A. Kramer, and S.M. Berget. 1998. A role for SRp54 during intron bridging of small introns with pyrimidine tracts upstream of the branch point. *Mol. Cell. Biol.* 18:5425–5434.

Komatsu, M., E. Kominami, K. Arahata, and T. Tsukahara. 1999. Cloning and characterization of two neural-salient serine/arginine-rich (NSSR) proteins involved in the regulation of alternative splicing in neurones. *Genes Cells.* 4:593–606.

Lasko, P. 2000. The *Drosophila* genome: translation factors and RNA binding proteins *J. Cell Biol.* 150:F51–F56.

Lo, P.C.H., and S.M. Mount. 1990. *Drosophila melanogaster* genes for U1 snRNA variants and their expression during development. *Nucleic Acids Res.* 18:6971–6979.

Lopez, J.A. 1998. Alternative Splicing of Pre-mRNA: Developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* 32:279–305.

Matunis, E.L., M.J. Matunis, and G. Dreyfuss. 1992. Characterization of the major hnRNP proteins from *Drosophila melanogaster. J. Cell Biol.* 116:257–269.

McLean, M.R., and B.C. Rymond. 1998. Yeast pre-mRNA splicing requires a pair of U1 snRNP-associated tetratricopeptide repeat proteins. *Mol. Cell. Biol.* 18:353–360.

Minvielle-Sebastia, L., and W. Keller. 1999. mRNA polyadenylation and its coupling to other RNA processing reactions and to transcription. *Curr. Opin. Cell Biol.* 11:352–357.

Mount, S.M. 1997. Genetic depletion reveals an essential role for an SR protein splicing factor in vertebrate cells. *Bioessays.* 19:189–192.

Polycarpou-Schwarz, M., S.I. Gunderson, S. Kandels-Lewis, B. Seraphin, and I.W. Mattaj. 1996. *Drosophila* SNF/D25 combines the functions of the two snRNP proteins U1A and U2B′′ that are encoded separately in human, potato and yeast. *RNA*. 2:11–23.

Puig, O., A. Gottschalk, P. Fabrizio, and B. Seraphin. 1999. Interaction of the U1 snRNP with nonconserved intronic sequences affects 5′ splice site selection. *Genes Dev*. 13:569–580.

Rigaut, G., A. Schevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. 1999. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* 17:1030–1032.

Saba, J.A., H. Busch, D. Wright, and R. Reddy. 1986. Isolation and characterization of two putative full-length *Drosophila* U4 small nuclear RNA genes. *J. Biol. Chem.* 261:8750–8753.

Saluz, H., R. Dudler, T. Schmidt, and E. Kubli. 1988. The localization and estimated copy number of *Drosophila melanogaster* U1, U4, U5 and U6 snRNA genes. *Nucleic Acids Res.* 16:3582.

Staley, J.P., and C. Guthrie. 1998. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell.* 92:315–326.

Stitzinger, S.M., T.R. Conrad, A.M. Zachlin, and H.K. Salz. 1999. Functional analysis of SNF, the *Drosophila* U1A/U2B′′ homolog: identification of dispensable and indispensable motifs for both snRNP assembly and function in vivo. *RNA*. 5:1440–1450.

Stojdl, D.F., and J.C. Bell. 1999. SR protein kinases: the splice of life. *Biochem. Cell Biol.* 77:293–298.

Tacke, R., and J.L. Manley. 1999. Determinants of SR protein specificity. *Curr. Opin. Cell Biol.* 11:358–362.

Tarn, W.Y., and J.A. Steitz. 1994. SR proteins can compensate for the loss of U1 snRNP functions in vitro. *Genes Dev.* 8:2704–2717.

Tarn, W.Y., and J.A. Steitz. 1996. A novel spliceosome containing U11, U12 and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell.* 84:801–811.

Tatusova, T.A., and T.L. Madden. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS (Fed. Eur. Microbiol. Soc.) Microbiol. Lett.* 174:247–250.

Tharun, S., W. He, A.E. Mayes, P. Lennertz, J.D. Beggs, R. Parker. 2000. Yeast Sm-like proteins function in mRNA decapping and decay. *Nature.* 404:515–518.

Vorbruggen, G., S. Onel, and H. Jackle. 2000. Localized expression of the *Drosophila* gene Dxl6, a novel member of the serine/arginine rich (SR) family of splicing factors. *Mech. Dev.* 90:309–312.

Will, C.L., C. Schneider, R. Reed, and R. Luhrmann. 1999. Identification of both shared and distinct proteins in the major and minor spliceosomes. *Science.* 284:2003–2005.

Yeakley, J.M, H. Tronchere, J. Olesen, J.A. Dyck, H.Y. Wang, and X.D. Fu. 1999. *J. Cell Biol.* 145:447–455.

Zahler, A.M., W.S. Lane, J.A. Stolk, and M.B. Roth. 1992. SR proteins: a conserved family of pre-mRNA splicing factors. *Genes Dev.* 6:837–847.

Zhang, D., and M. Rosbash. 1999. Identification of eight proteins that cross-link to pre-mRNA in the yeast commitment complex. *Genes Dev.* 13:581–592.